Building an ETL Pipeline using Azure Data Services

DESCRIPTION

Use the data analytics stack to build a data pipeline using Data Factory, Databricks and Synapse.

Problem Statement:

As a Data Engineer, you've been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will be collecting the data from SQL Database using ADF and the transformed data will be the source for databricks to run complex transformations and once data is analysed using Databricks, it is synced into synapse analytics data warehouse as historical dataset for enabling various analytics.

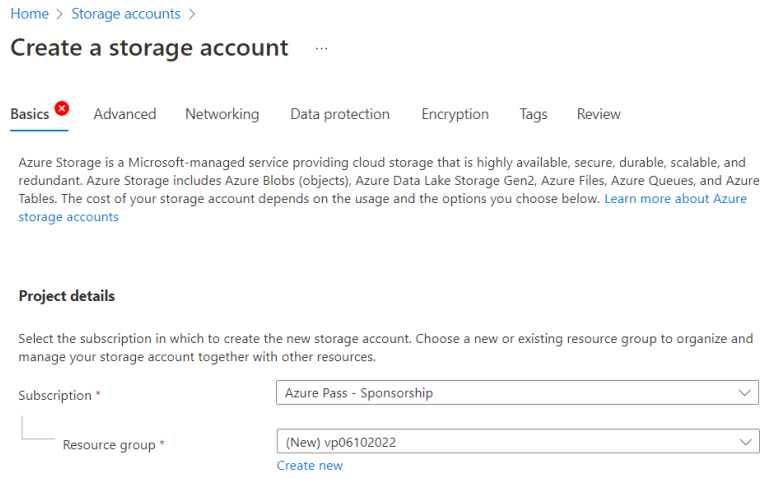
Domain: Analytics

Steps for building ETL pipeline :

In this project, perform the following steps:

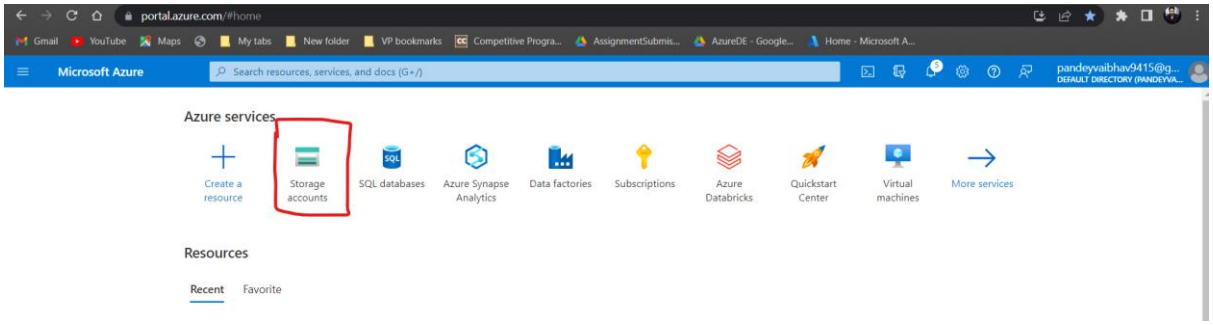
● Create a Resource Group.

To create a resource group we need to go to Create Storage account then fill the details to make a Resource group.

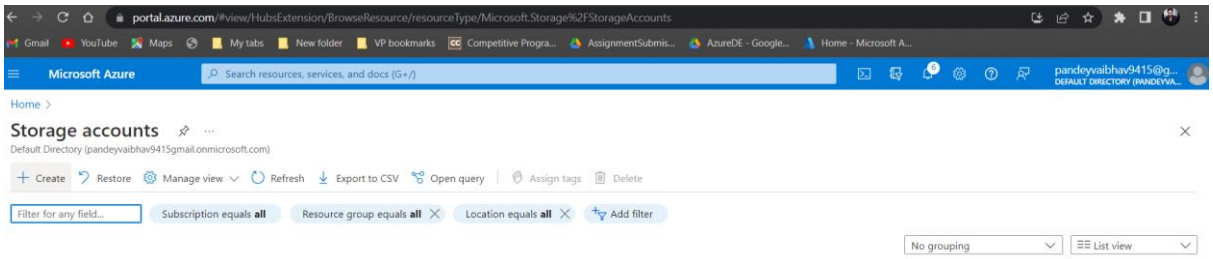


● Create a Storage account.

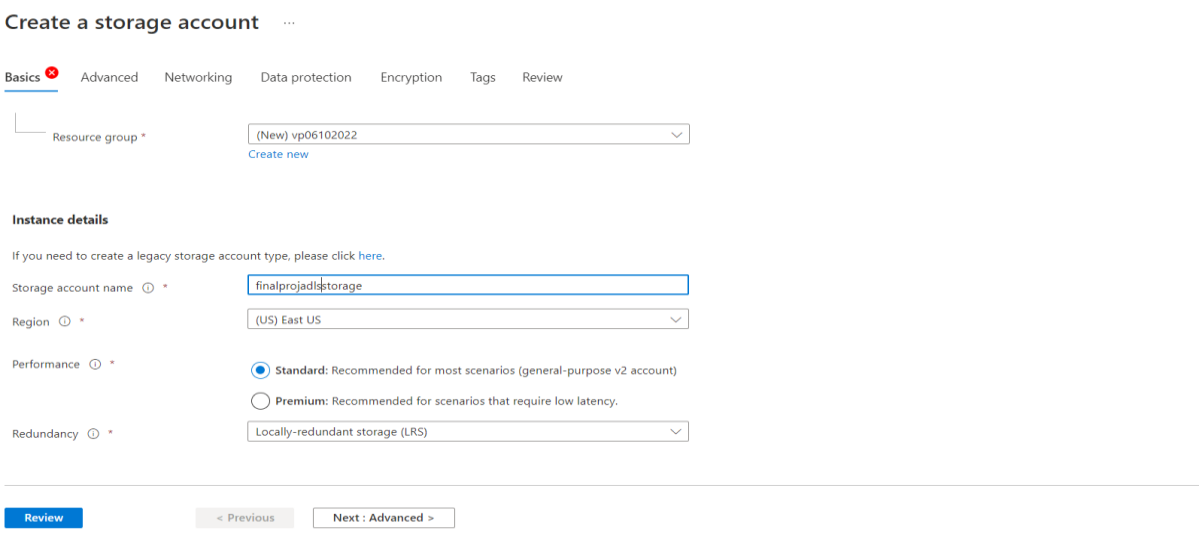
Step 1 : Open Storage account in azure portal @portal.azure.com/#home

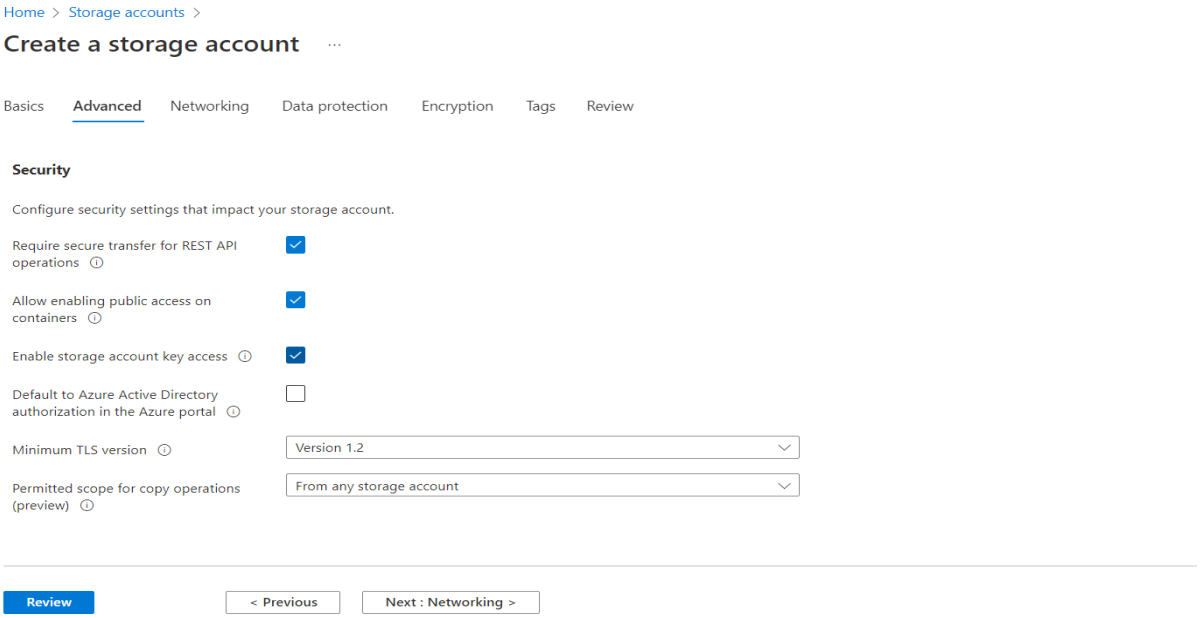


Step 2 : Click on Create option to create a new storage account

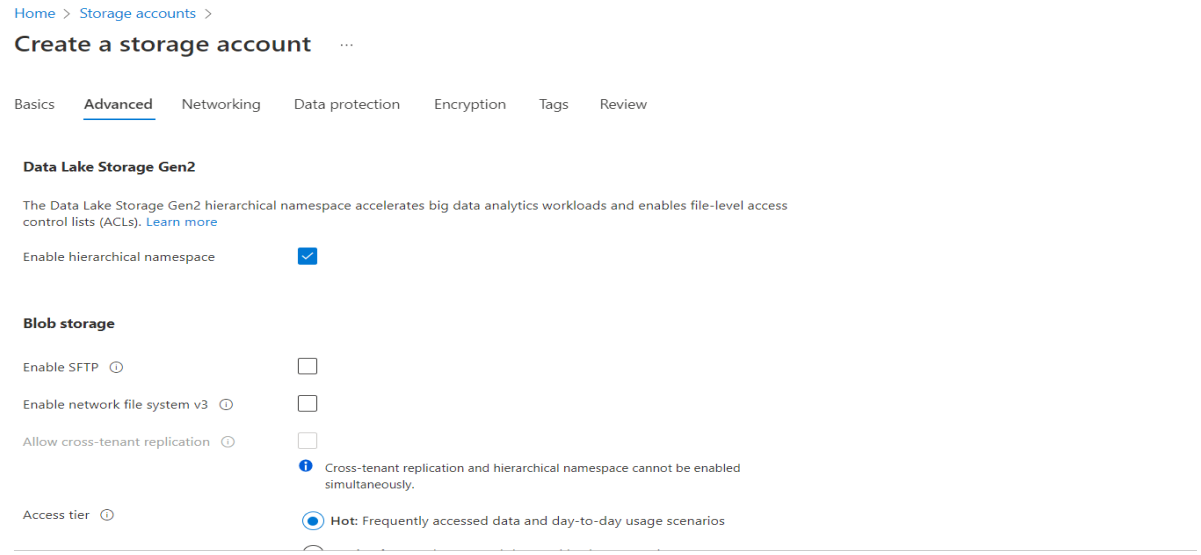


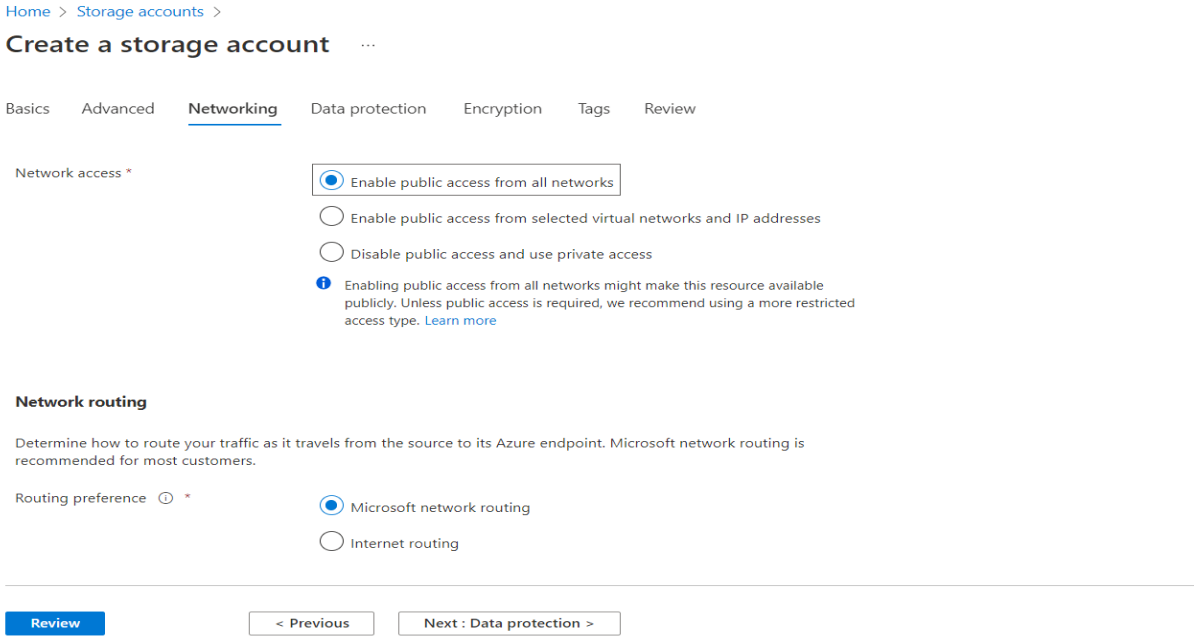
Select redundancy as ‘LRS’ and Performance as Standard

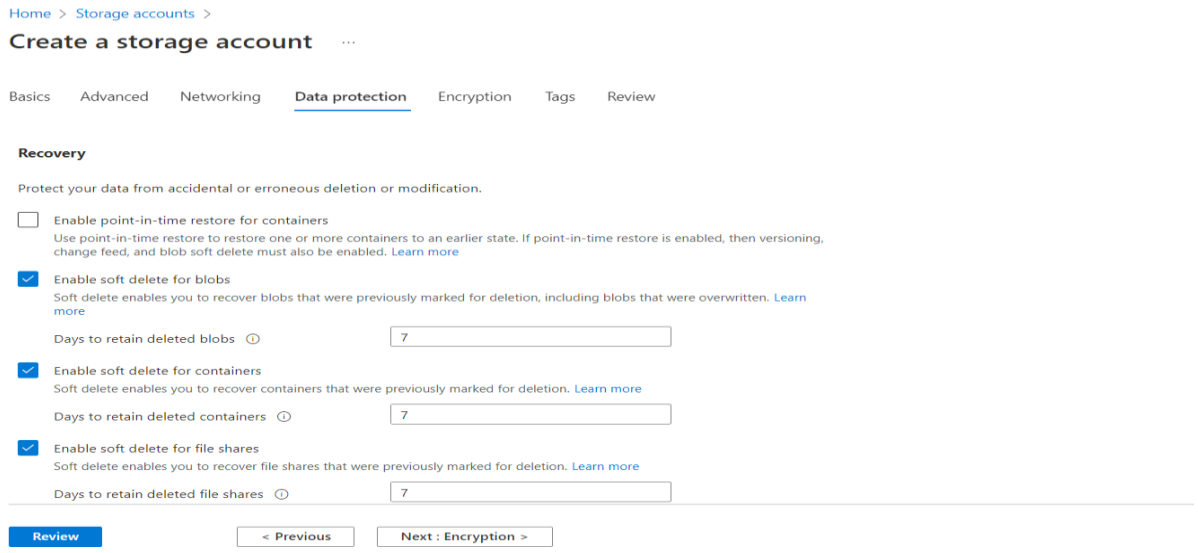




Check the “Enable Hierarchical space” box to ensure the account to be created as ADLSv2 storage.







Keep “Encryption ” and “Tags” as Default . Then click on Create.

Home > Storage accounts >

Create a storage account ...

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review

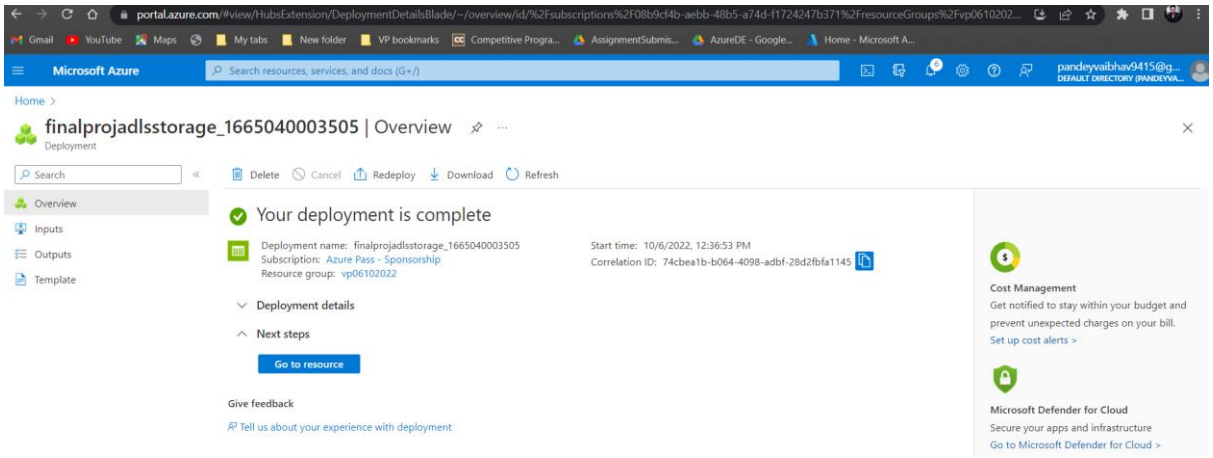
Basics

Subscription	Azure Pass - Sponsorship
Resource Group	vp06102022
Location	eastus
Storage account name	finalprojadlsstorage
Deployment model	Resource manager
Performance	Standard
Replication	Locally-redundant storage (LRS)

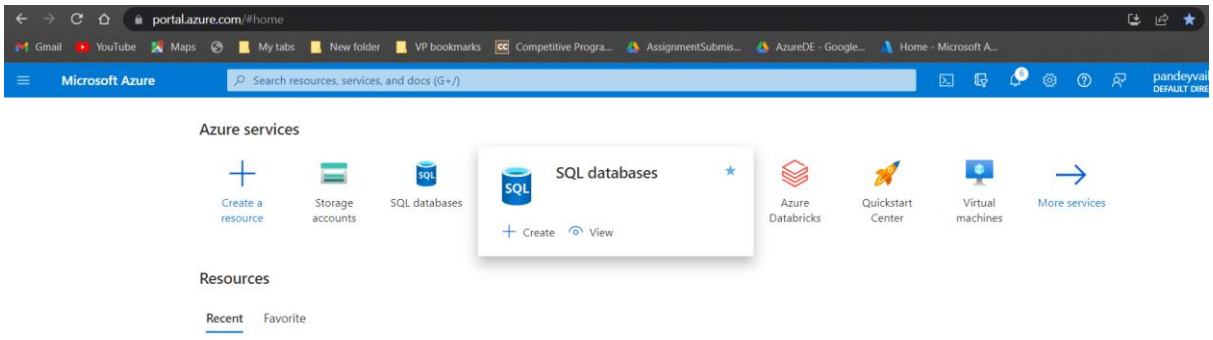
Advanced

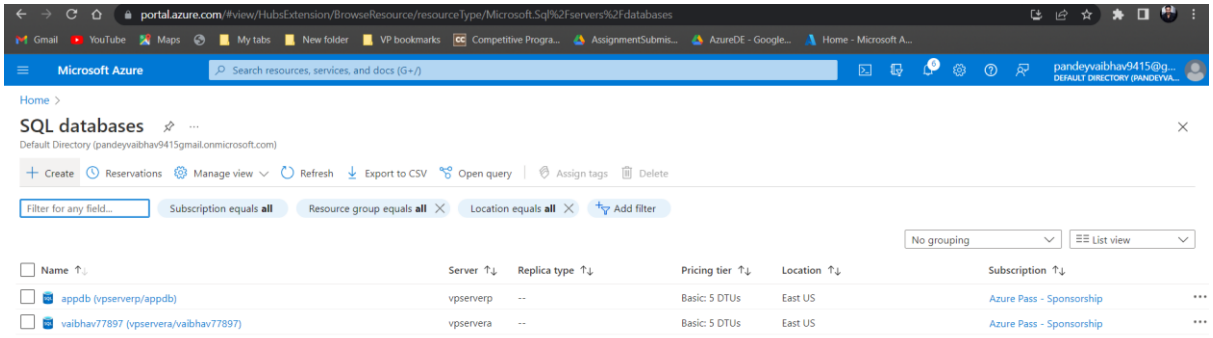
Secure transfer	Enabled
Allow storage account key access	Enabled
Allow cross-tenant replication	Disabled
Default to Azure Active Directory authorization in the Azure portal	Disabled
Blob public access	Enabled

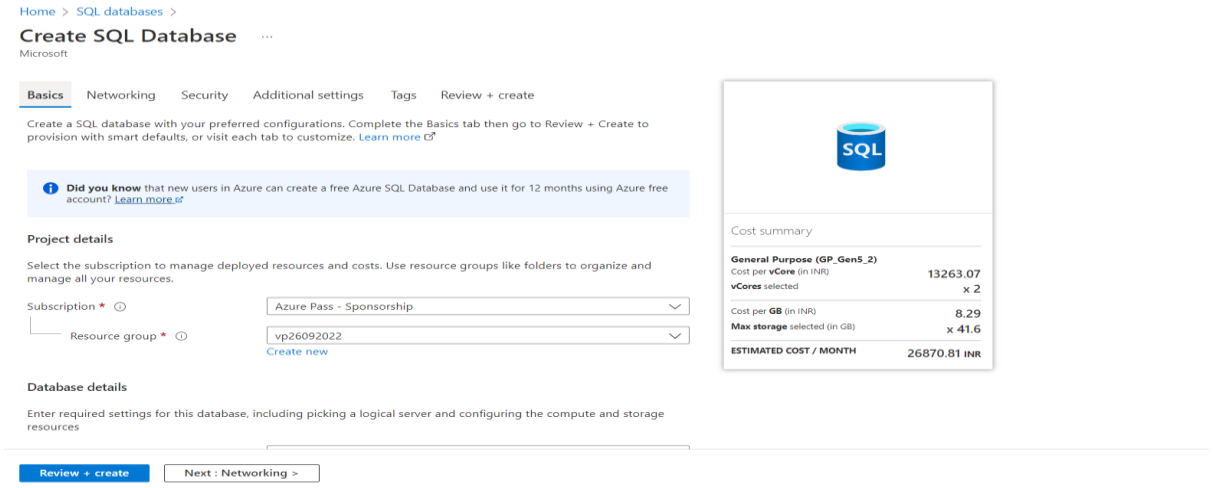
Create | < Previous | Next > | Download a template for automation

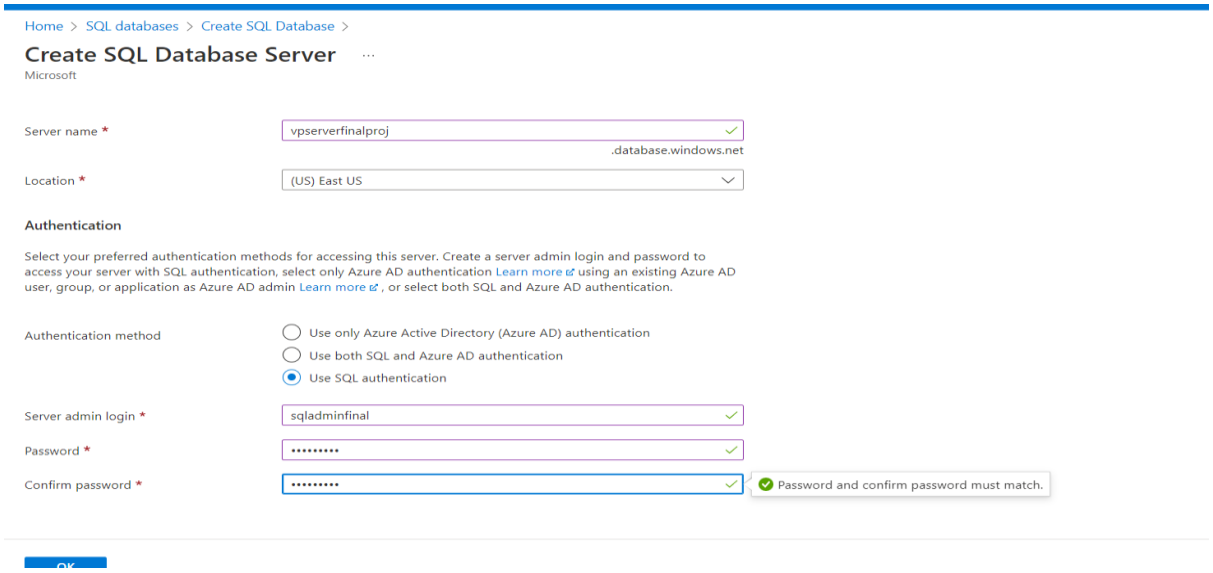


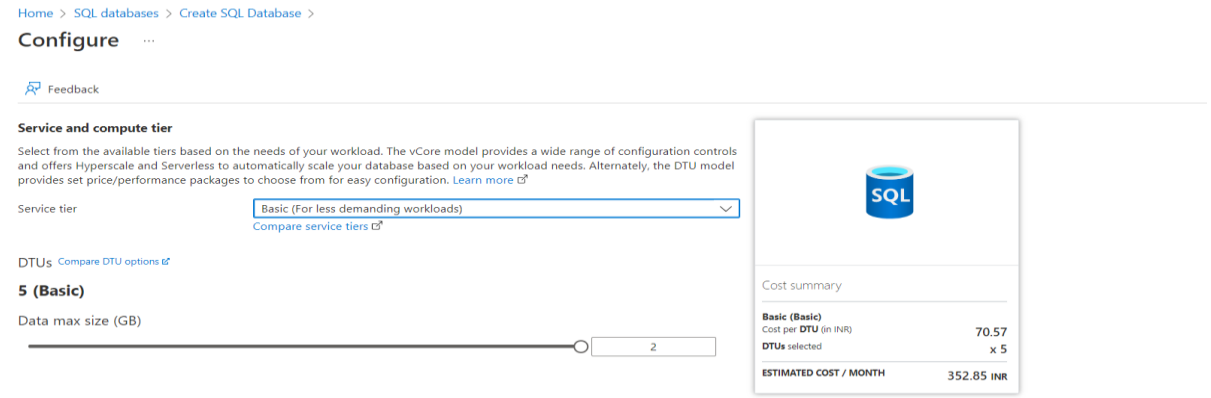
● Create an Azure SQL Database.



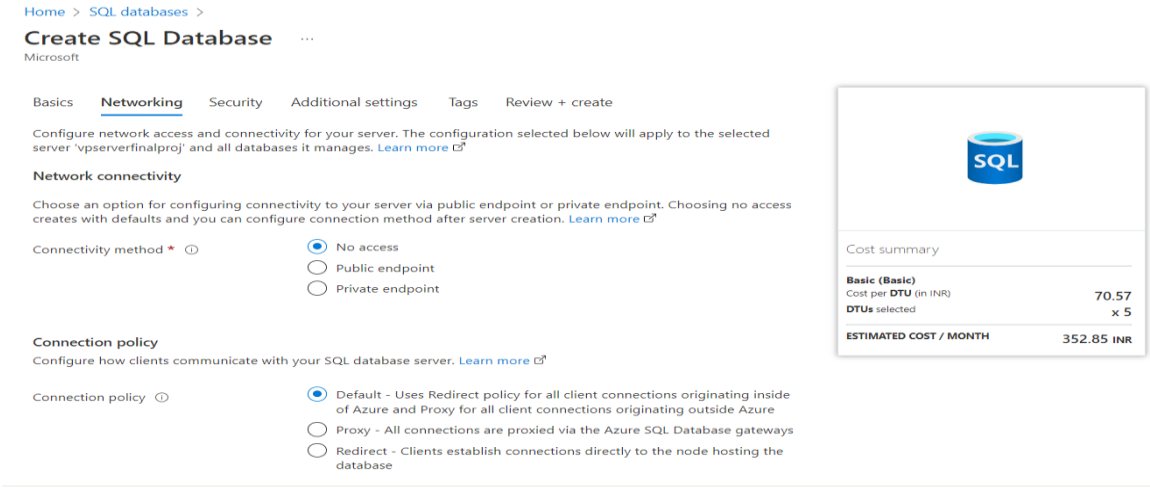

portal.azure.com/#view/HubsExtension/BrowseResource/resourceType/Microsoft.Sql%2Fservers%2Fdatabases
Microsoft Azure
Search resources, services, and docs (G+/)
Home >
SQL databases
Default Directory (pandeyvaibhav9415gmail.onmicrosoft.com)
Create
Reservations
Manage view
Refresh
Export to CSV
Open query
Assign tags
Delete
Filter for any field...
Subscription equals all
Resource group equals all
Location equals all
Add filter
No grouping
List view
Name
Server
Replica type
Pricing tier
Location
Subscription
appdb (vpserverp/appdb)
vpserverp
--
Basic: 5 DTUs
East US
Azure Pass - Sponsorship
vaibhav77897 (vpservera/vaibhav77897)
vpservera
--
Basic: 5 DTUs
East US
Azure Pass - Sponsorship

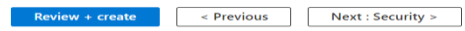

Home > SQL databases >
Create SQL Database
Microsoft
Basics
Networking
Security
Additional settings
Tags
Review + create
Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. Learn more
Did you know that new users in Azure can create a free Azure SQL Database and use it for 12 months using Azure free account? Learn more
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 Pass - Sponsorship
Resource group *
vp26092022
Create new
Database details
Enter required settings for this database, including picking a logical server and configuring the compute and storage resources
SQL
Cost summary
General Purpose (GP_Gen5_2)
Cost per vCore (in INR)
13263.07
vCores selected
x 2
Cost per GB (in INR)
8.29
Max storage selected (in GB)
x 41.6
ESTIMATED COST / MONTH
26870.81 INR
Review + create
Next : Networking >

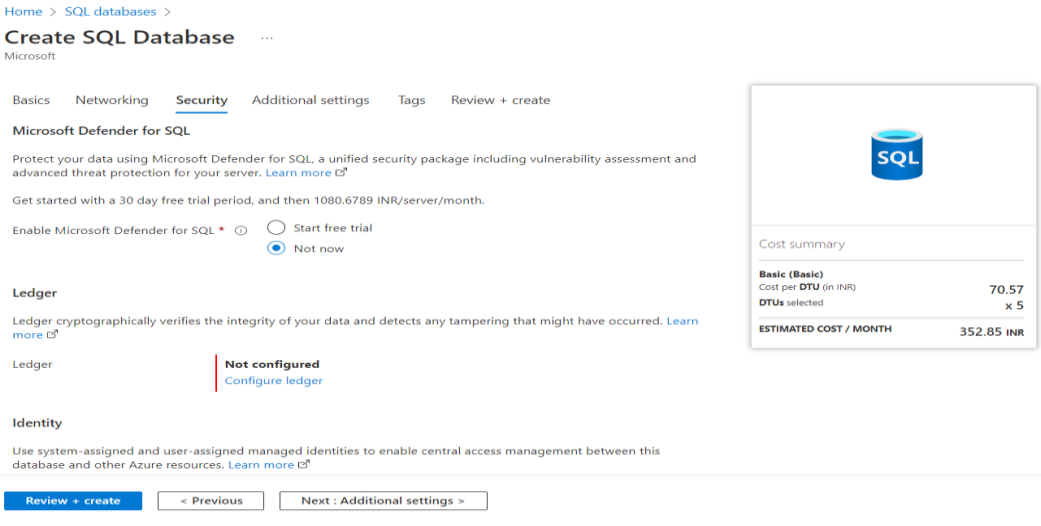

Home > SQL databases > Create SQL Database >
Create SQL Database Server
Microsoft
Server name *
vpserverfinalproj
.database.windows.net
Location *
(US) East US
Authentication
Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Azure AD authentication Learn more using an existing Azure AD user, group, or application as Azure AD admin Learn more, or select both SQL and Azure AD authentication.
Authentication method
Use only Azure Active Directory (Azure AD) authentication
Use both SQL and Azure AD authentication
Use SQL authentication
Server admin login *
sqladminfinal
Password *
Confirm password *
Password and confirm password must match.
OK


Home > SQL databases > Create SQL Database >
Configure
Feedback
Service and compute tier
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages to choose from for easy configuration. Learn more
Service tier
Basic (For less demanding workloads)
Compare service tiers
DTUs Compare DTU options
5 (Basic)
Data max size (GB)
2
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR

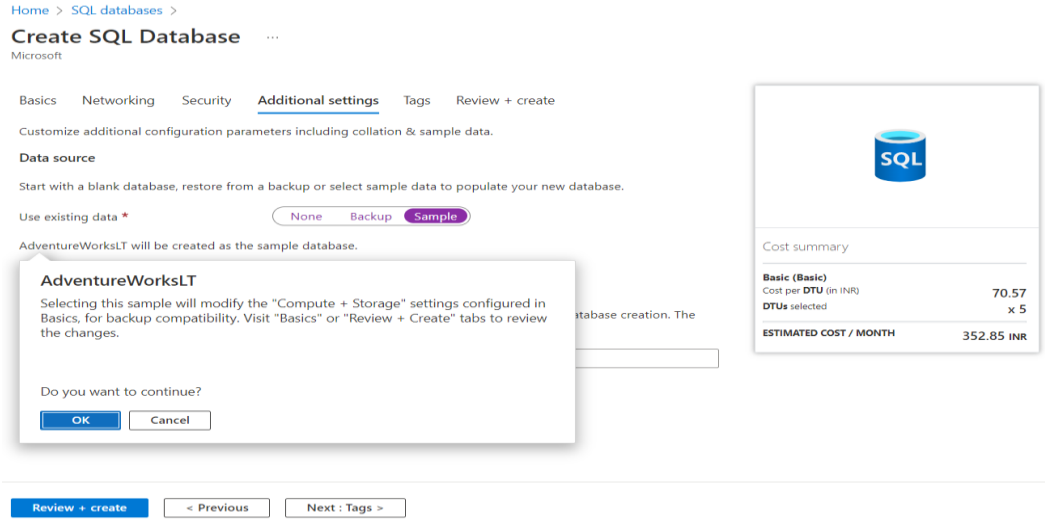

Apply


Home > SQL databases >
Create SQL Database
Microsoft
Basics
Networking
Security
Additional settings
Tags
Review + create
Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'vpserverfinalproj' and all databases it manages. Learn more
Network connectivity
Choose an option for configuring connectivity to your server via public endpoint or private endpoint. Choosing no access creates with defaults and you can configure connection method after server creation. Learn more
Connectivity method *
No access
Public endpoint
Private endpoint
Connection policy
Configure how clients communicate with your SQL database server. Learn more
Connection policy
Default - Uses Redirect policy for all client connections originating inside of Azure and Proxy for all client connections originating outside Azure
Proxy - All connections are proxied via the Azure SQL Database gateways
Redirect - Clients establish connections directly to the node hosting the database
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR

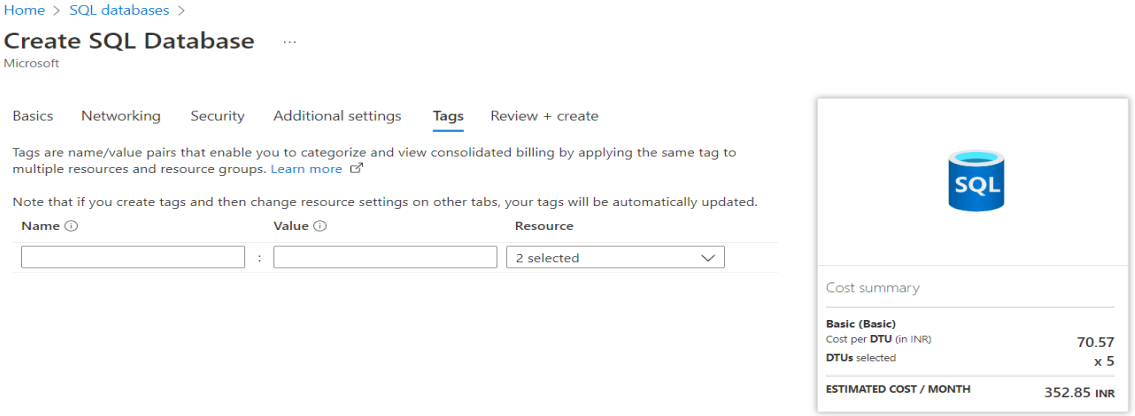

Review + create
< Previous
Next : Security >

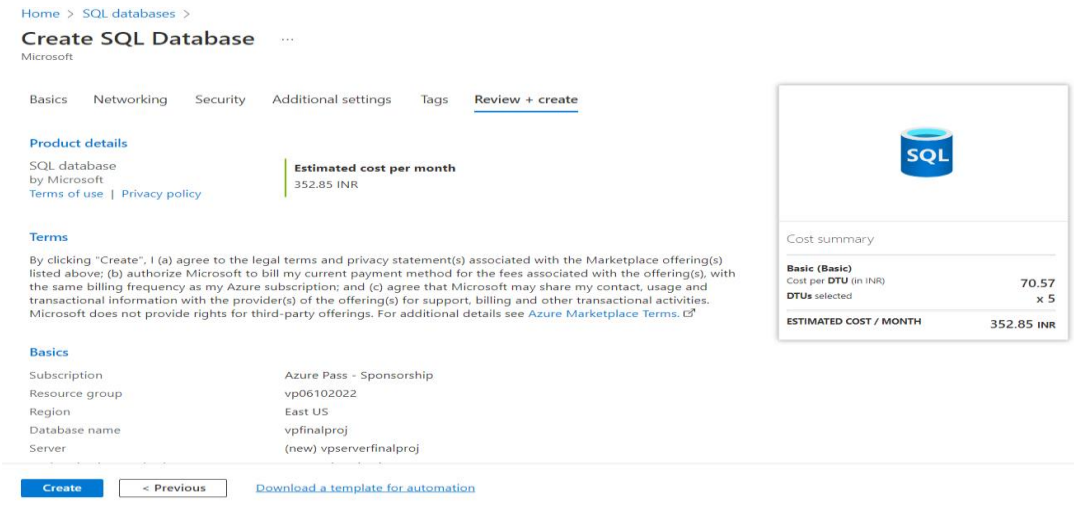

Home > SQL databases >
Create SQL Database
Microsoft
Basics
Networking
Security
Additional settings
Tags
Review + create
Microsoft Defender for SQL
Protect your data using Microsoft Defender for SQL, a unified security package including vulnerability assessment and advanced threat protection for your server. Learn more
Get started with a 30 day free trial period, and then 1080.6789 INR/server/month.
Enable Microsoft Defender for SQL *
Start free trial
Not now
Ledger
Ledger cryptographically verifies the integrity of your data and detects any tampering that might have occurred. Learn more
Ledger
Not configured
Configure ledger
Identity
Use system-assigned and user-assigned managed identities to enable central access management between this database and other Azure resources. Learn more
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR

Review + create
< Previous
Next : Additional settings >

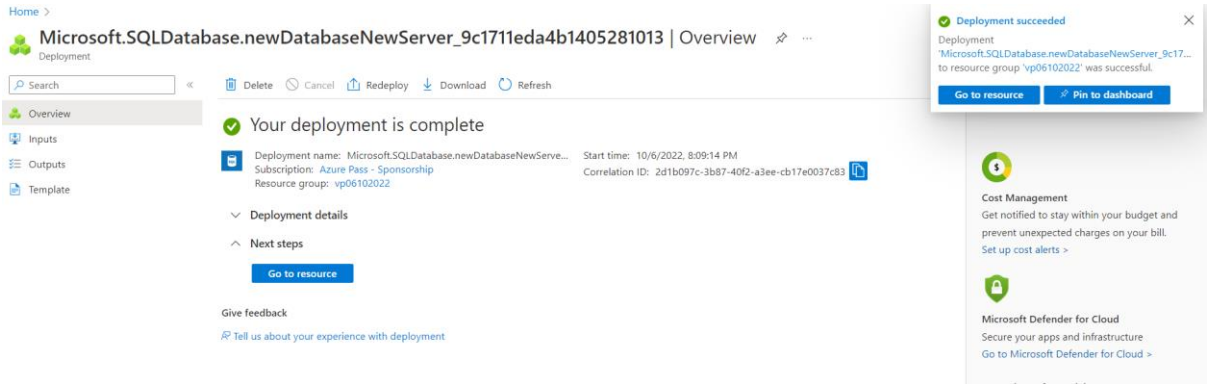

Home > SQL databases >
Create SQL Database
Microsoft
Basics Networking Security Additional settings Tags Review + create
Customize additional configuration parameters including collation & sample data.
Data source
Start with a blank database, restore from a backup or select sample data to populate your new database.
Use existing data *
None Backup Sample
AdventureWorksLT will be created as the sample database.
AdventureWorksLT
Selecting this sample will modify the "Compute + Storage" settings configured in Basics, for backup compatibility. Visit "Basics" or "Review + Create" tabs to review the changes.
Do you want to continue?
OK
Cancel
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR

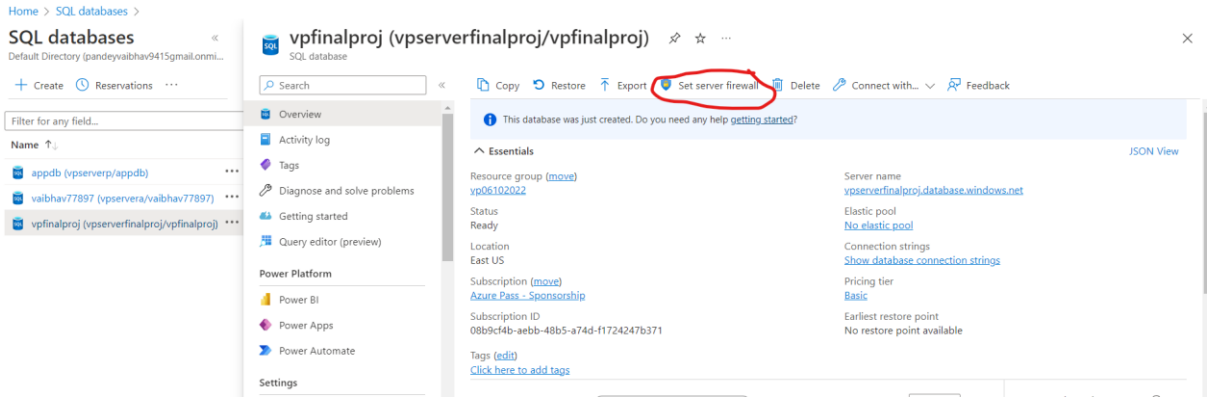
Review + create
< Previous
Next : Tags >

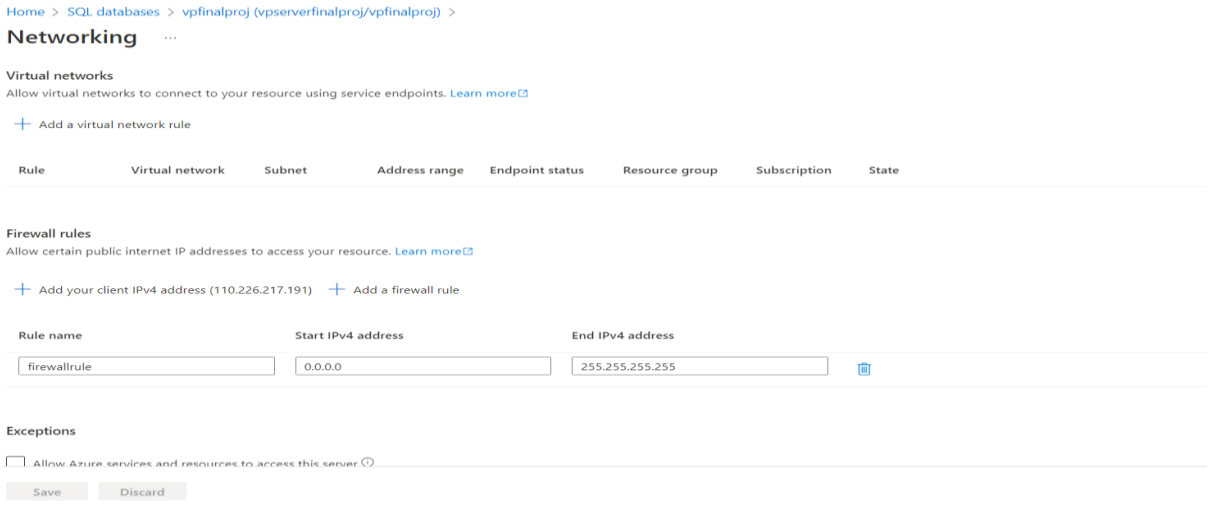

Home > SQL databases >
Create SQL Database
Microsoft
Basics Networking Security Additional settings Tags Review + create
Tags are name/value pairs that enable you to categorize and view consolidated billing by applying the same tag to multiple resources and resource groups. Learn more
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
2 selected
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR


Home > SQL databases >
Create SQL Database
Microsoft
Basics Networking Security Additional settings Tags Review + create
Product details
SQL database
by Microsoft
Terms of use | Privacy policy
Estimated cost per month
352.85 INR
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 Pass - Sponsorship
Resource group
vp06102022
Region
East US
Database name
vpfinalproj
Server
(new) vpserverfinalproj
Cost summary
Basic (Basic)
Cost per DTU (in INR)
70.57
DTUs selected
x 5
ESTIMATED COST / MONTH
352.85 INR

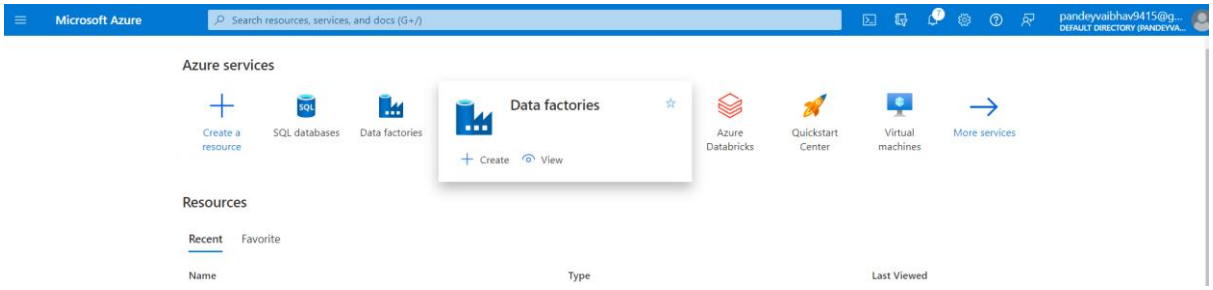
Create
< Previous
Download a template for automation

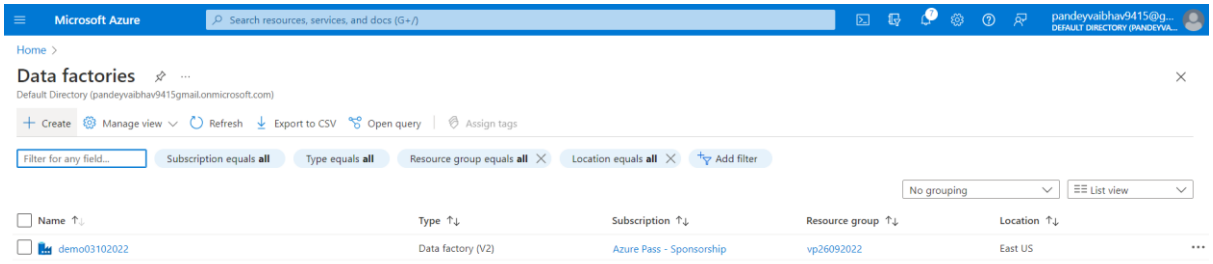

Home >
Microsoft.SQLDatabase.newDatabaseNewServer_9c1711eda4b1405281013 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name: Microsoft.SQLDatabase.newDatabaseNewServe...
Subscription: Azure Pass - Sponsorship
Resource group: vp06102022
Start time: 10/6/2022, 8:09:14 PM
Correlation ID: 2d1b097c-3b87-40f2-a3ee-cb17e0037c83
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment
Deployment succeeded
Deployment 'Microsoft.SQLDatabase.newDatabaseNewServer_9c17... to resource group 'vp06102022' was successful.
Go to resource
Pin to dashboard
Cost Management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

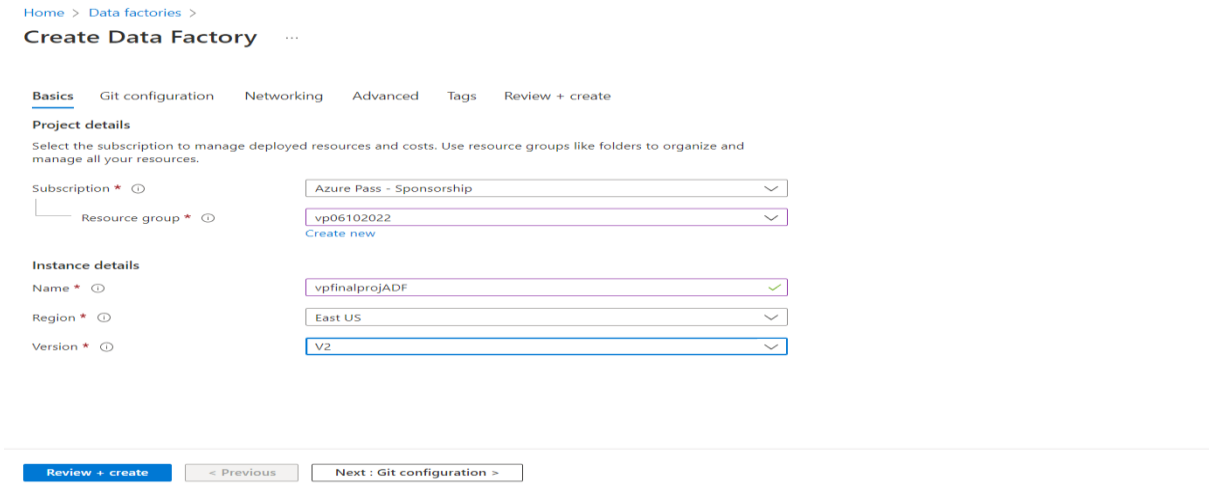

Home > SQL databases >
SQL databases
Default Directory (pandeyvaibhav9415gmail.onmi...
Create
Reservations
Filter for any field...
Name
appdb (vpserverp/appdb)
vaibhav77897 (vpservera/vaibhav77897)
vpfinalproj (vpserverfinalproj/vpfinalproj)
vpfinalproj (vpserverfinalproj/vpfinalproj)
SQL database
Search
Overview
Activity log
Tags
Diagnose and solve problems
Getting started
Query editor (preview)
Power Platform
Power BI
Power Apps
Power Automate
Settings
Copy
Restore
Export
Set server firewall
Delete
Connect with...
Feedback
This database was just created. Do you need any help getting started?
Essentials
JSON View
Resource group (move)
vp06102022
Status
Ready
Location
East US
Subscription (move)
Azure Pass - Sponsorship
Subscription ID
08b9cf4b-aebb-48b5-a74d-f1724247b371
Tags (edit)
Click here to add tags
Server name
vpserverfinalproj.database.windows.net
Elastic pool
No elastic pool
Connection strings
Show database connection strings
Pricing tier
Basic
Earliest restore point
No restore point available

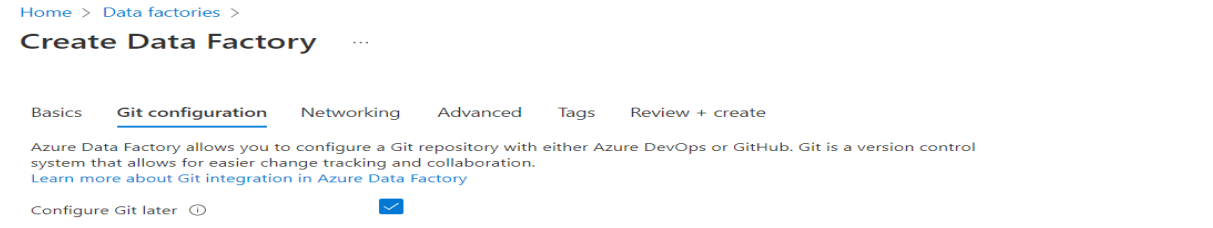

Home > SQL databases > vpfinalproj (vpserverfinalproj/vpfinalproj) >
Networking
Virtual networks
Allow virtual networks to connect to your resource using service endpoints. Learn more
Add a virtual network rule
Rule
Virtual network
Subnet
Address range
Endpoint status
Resource group
Subscription
State
Firewall rules
Allow certain public internet IP addresses to access your resource. Learn more
Add your client IPv4 address (110.226.217.191)
Add a firewall rule
Rule name
Start IPv4 address
End IPv4 address
firewallrule
0.0.0.0
255.255.255.255
Exceptions
Allow Azure services and resources to access this server
Save
Discard

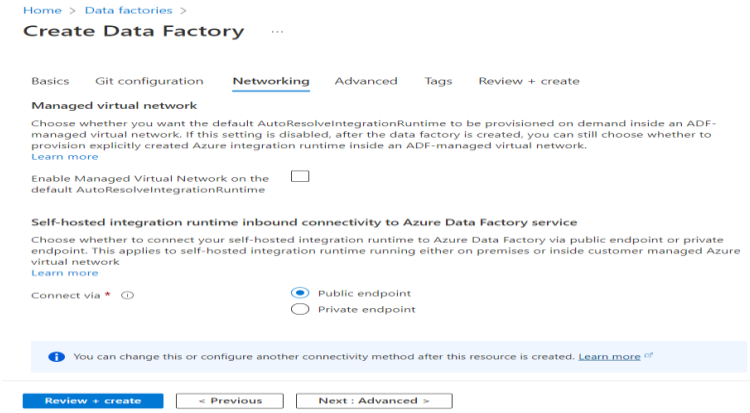
- **Create a data factory.**

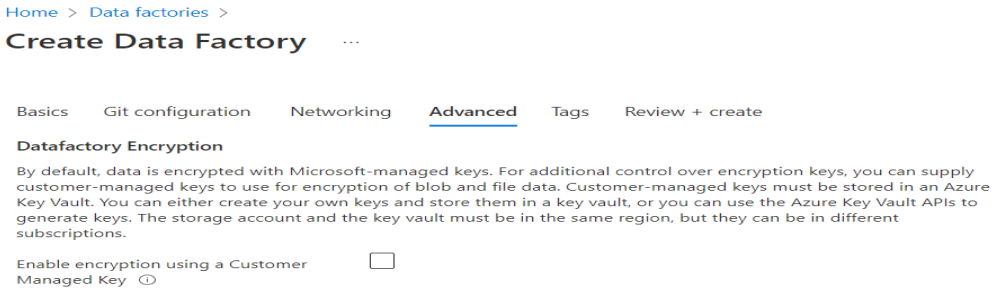


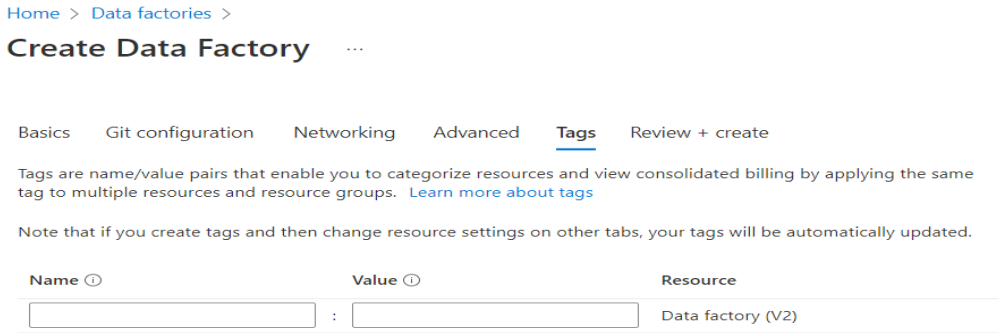







Home > Data factories >
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Managed virtual network
Choose whether you want the default AutoResolveIntegrationRuntime to be provisioned on demand inside an ADF-managed virtual network. If this setting is disabled, after the data factory is created, you can still choose whether to provision explicitly created Azure integration runtime inside an ADF-managed virtual network.
Learn more
Enable Managed Virtual Network on the default AutoResolveIntegrationRuntime
Self-hosted integration runtime inbound connectivity to Azure Data Factory service
Choose whether to connect your self-hosted integration runtime to Azure Data Factory via public endpoint or private endpoint. This applies to self-hosted integration runtime running either on premises or inside customer managed Azure virtual network
Learn more
Connect via *
Public endpoint
Private endpoint
You can change this or configure another connectivity method after this resource is created. Learn more
Review + create
< Previous
Next : Advanced >


Home > Data factories >
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Datafactory Encryption
By default, data is encrypted with Microsoft-managed keys. For additional control over encryption keys, you can supply customer-managed keys to use for encryption of blob and file data. Customer-managed keys must be stored in an Azure Key Vault. You can either create your own keys and store them in a key vault, or you can use the Azure Key Vault APIs to generate keys. The storage account and the key vault must be in the same region, but they can be in different subscriptions.
Enable encryption using a Customer Managed Key


Home > Data factories >
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. Learn more about tags
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
Data factory (V2)

Home > Data factories >

Create Data Factory ...

Validation Passed

Basics | Git configuration | Networking | Advanced | Tags | **Review + create**

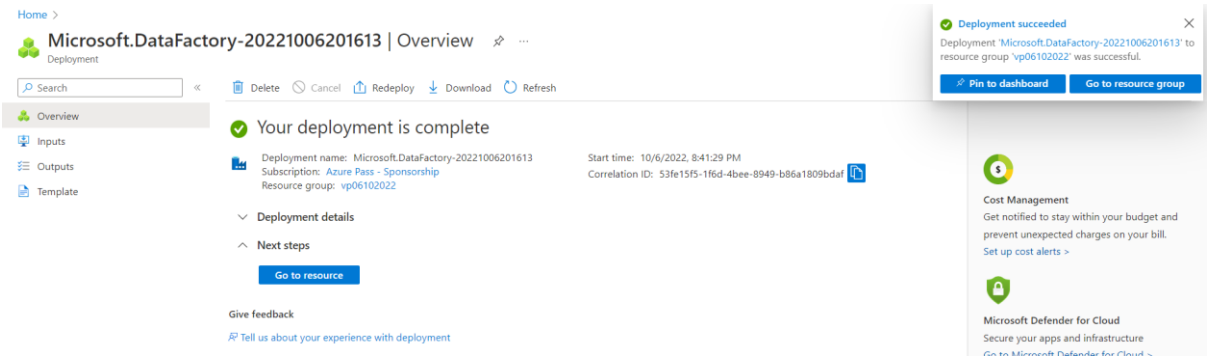
TERMS

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. See the Azure Marketplace Terms for additional details.

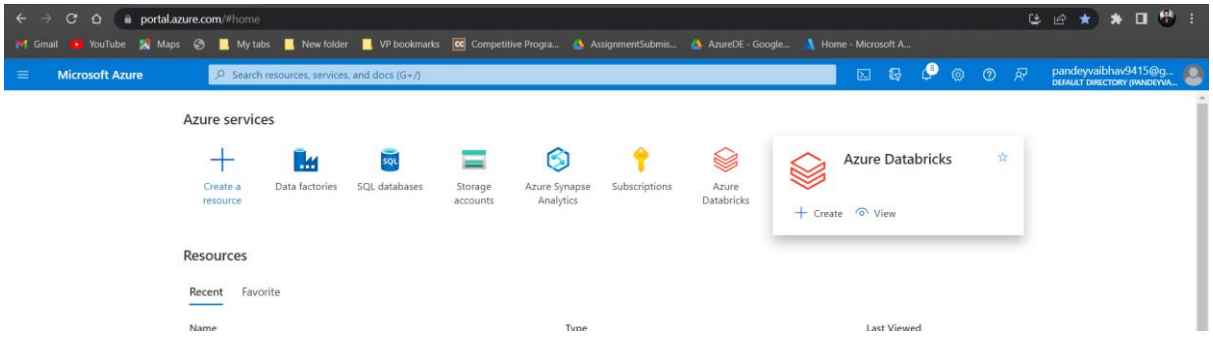
Basics

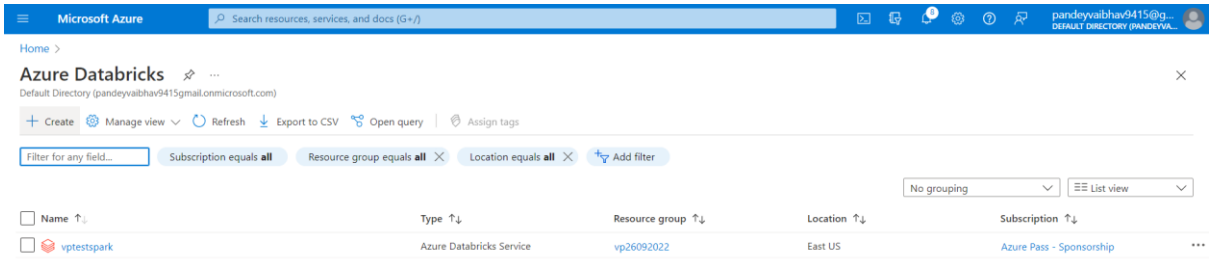
Subscription	Azure Pass - Sponsorship
Resource group	vp06102022
Name	vpfinalprojADF
Region	East US
Version	V2

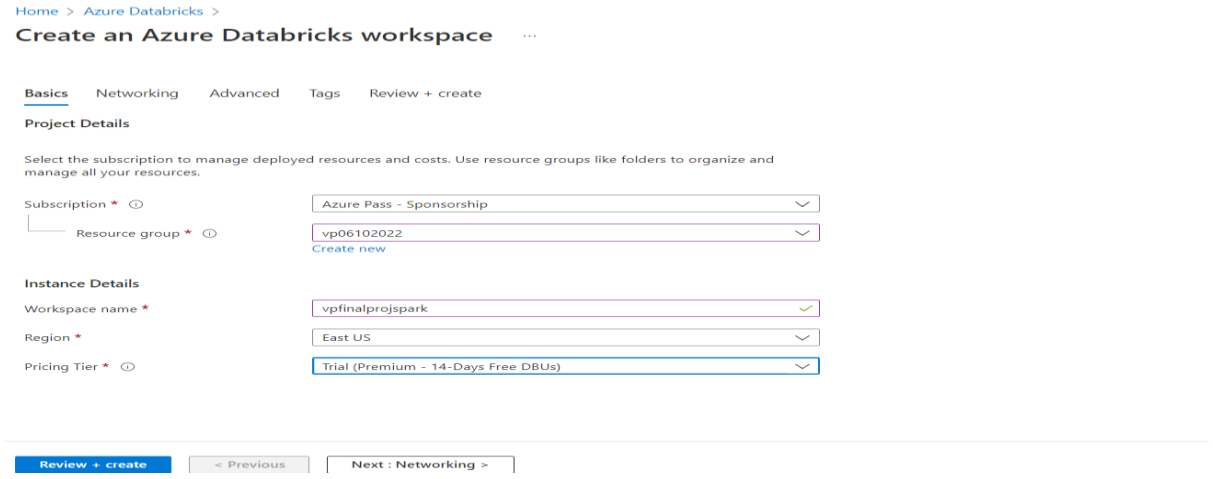
Create | < Previous | Next

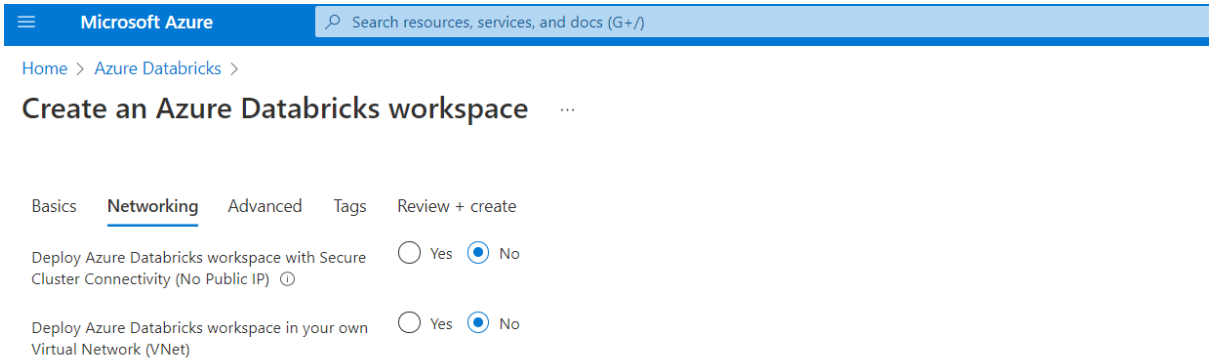


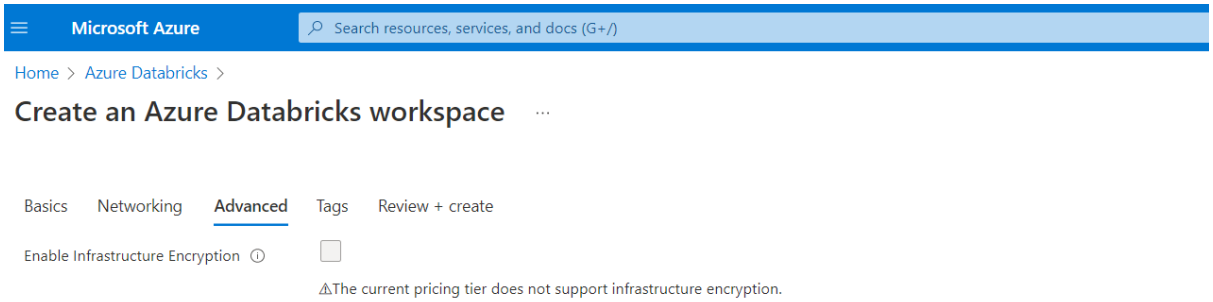
• Configure Databricks cluster

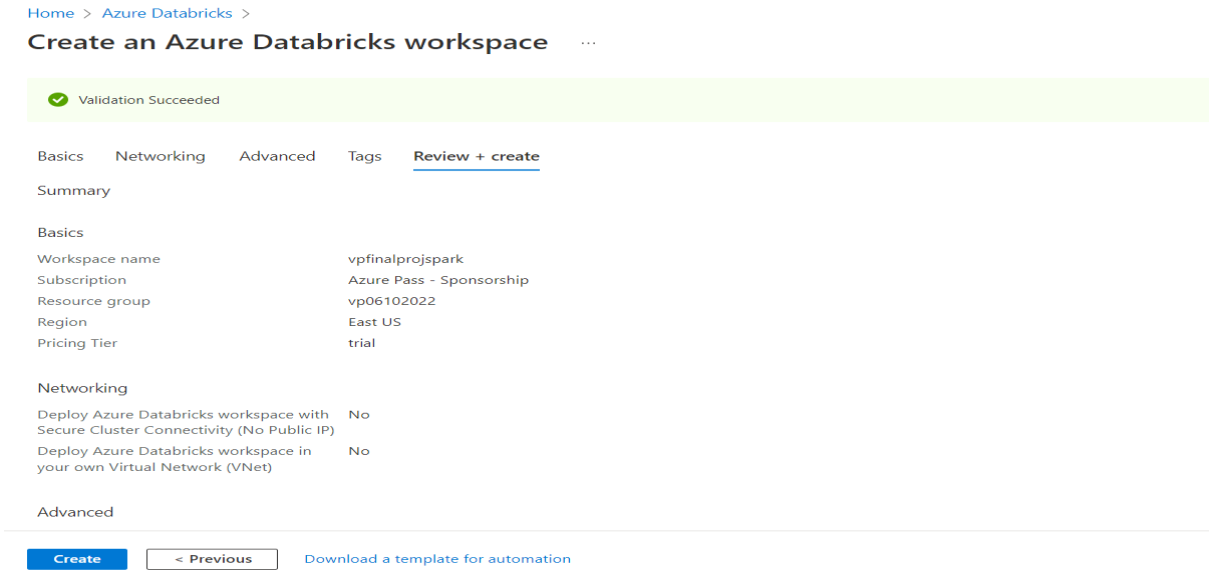


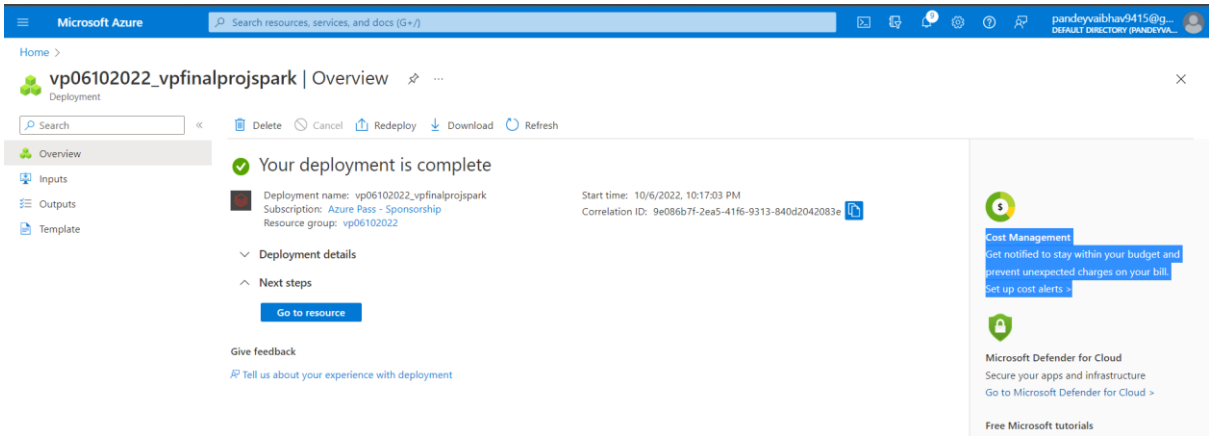




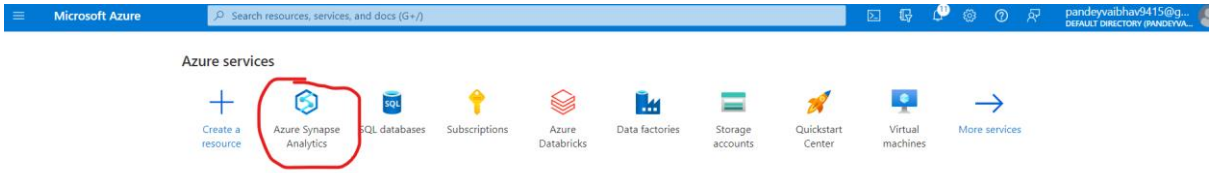


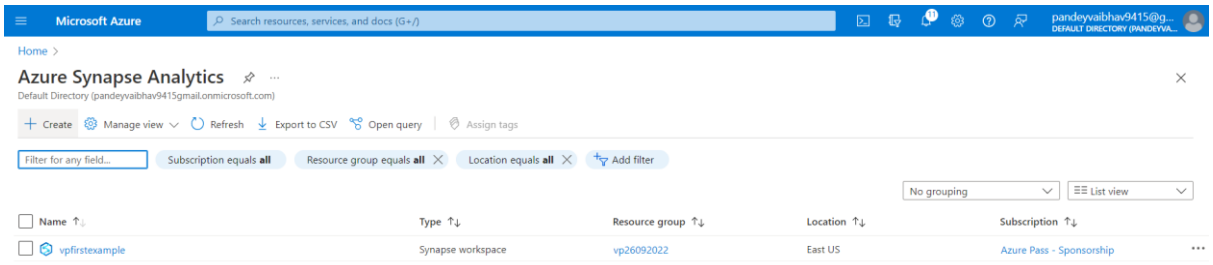

Microsoft Azure
Search resources, services, and docs (G+/)
Home > Azure Databricks >
Create an Azure Databricks workspace
Basics Networking Advanced Tags Review + create
Enable Infrastructure Encryption
⚠The current pricing tier does not support infrastructure encryption.

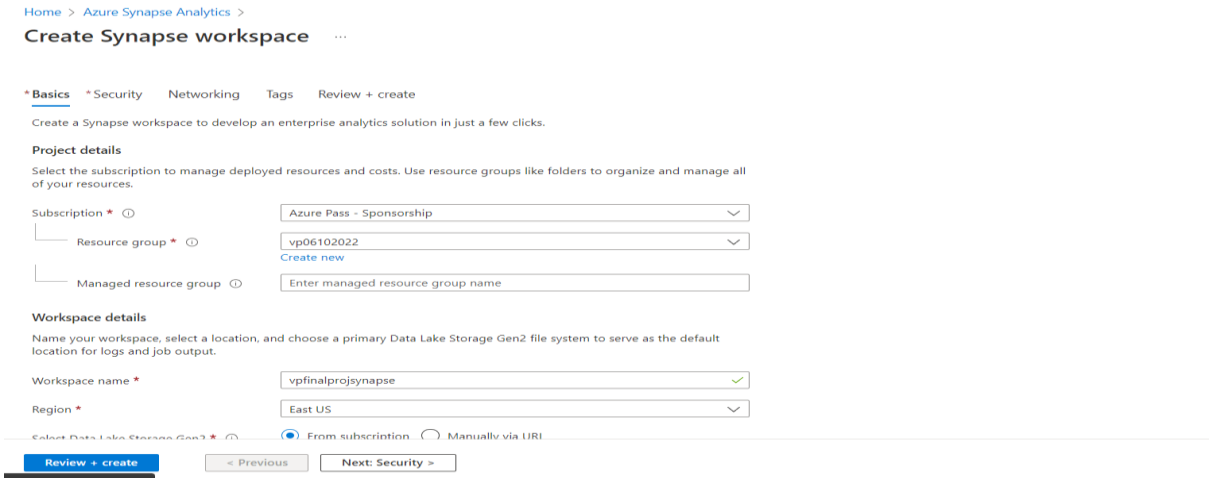

Home > Azure Databricks >
Create an Azure Databricks workspace
Validation Succeeded
Basics Networking Advanced Tags Review + create
Summary
Basics
Workspace name vpfinalprojspark
Subscription Azure Pass - Sponsorship
Resource group vp06102022
Region East US
Pricing Tier trial
Networking
Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) No
Deploy Azure Databricks workspace in your own Virtual Network (VNet) No
Advanced
Create
< Previous
Download a template for automation

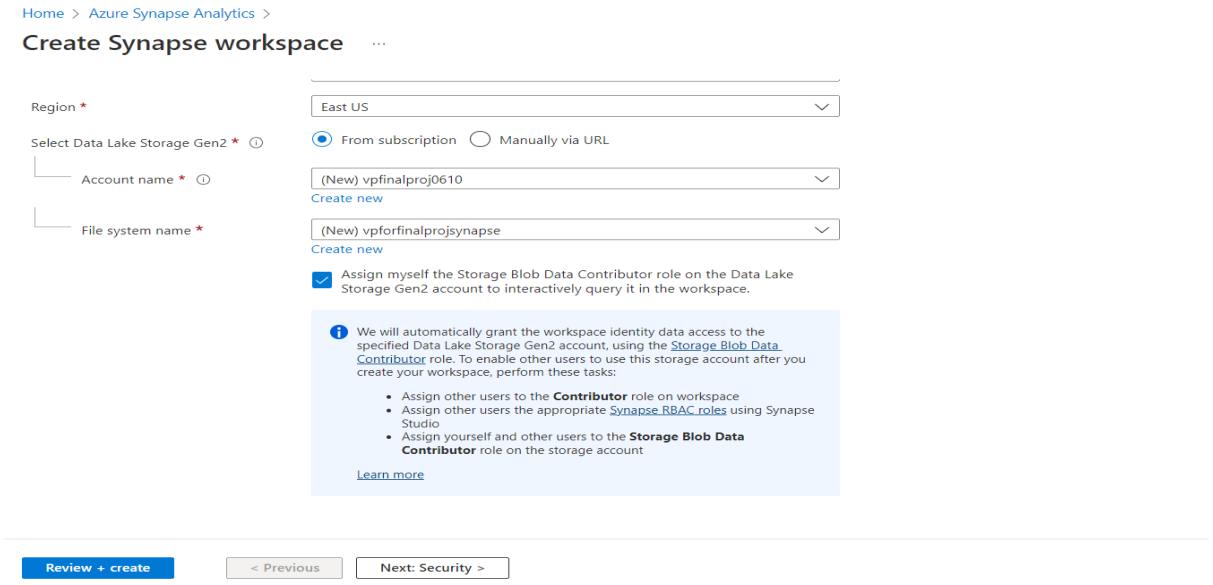

Microsoft Azure
Search resources, services, and docs (G+/)
pandeyvaibhav9415@g...
DEFAULT DIRECTORY (PANDEYVA...
Home >
vp06102022_vpfinalprojspark | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete Cancel Redeploy Download Refresh
Your deployment is complete
Deployment name: vp06102022_vpfinalprojspark
Subscription: Azure Pass - Sponsorship
Resource group: vp06102022
Start time: 10/6/2022, 10:17:03 PM
Correlation ID: 9e086b7f-2ea5-41f6-9313-840d2042083e
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 >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials

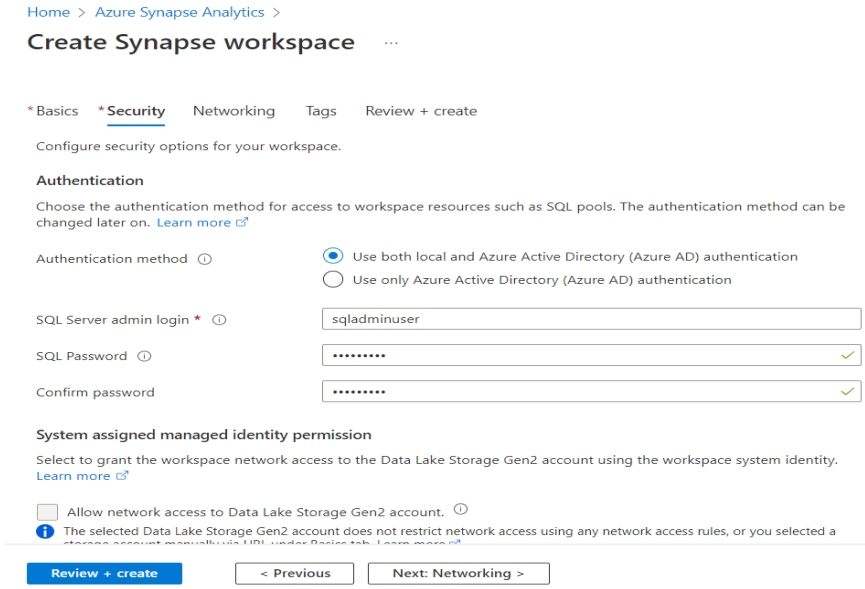
• Create Synapse analytics Data Warehouse.



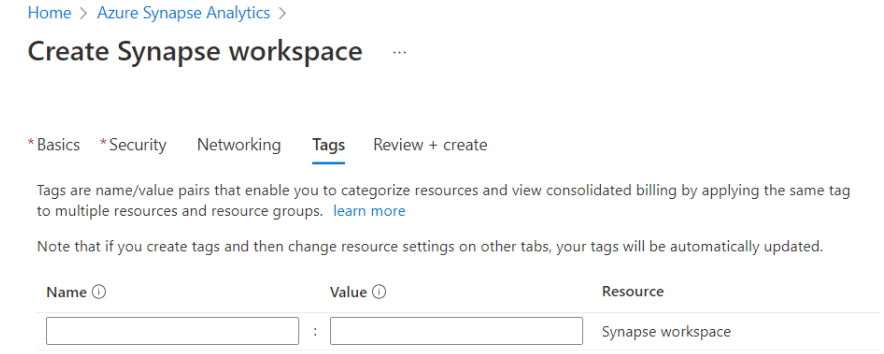


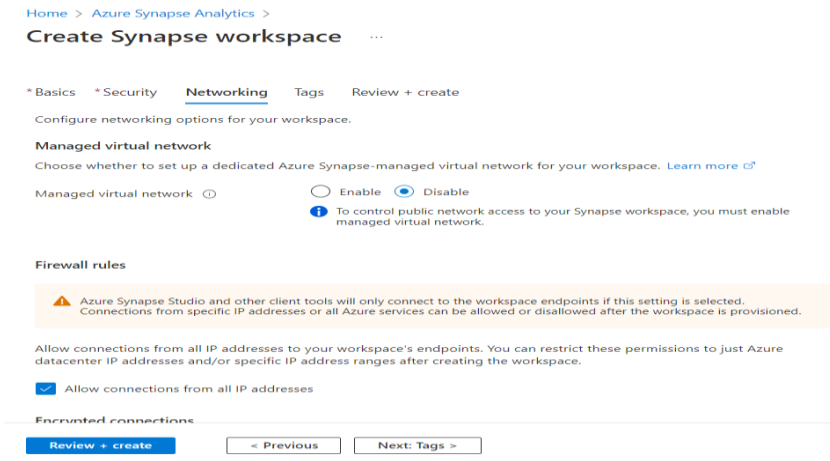


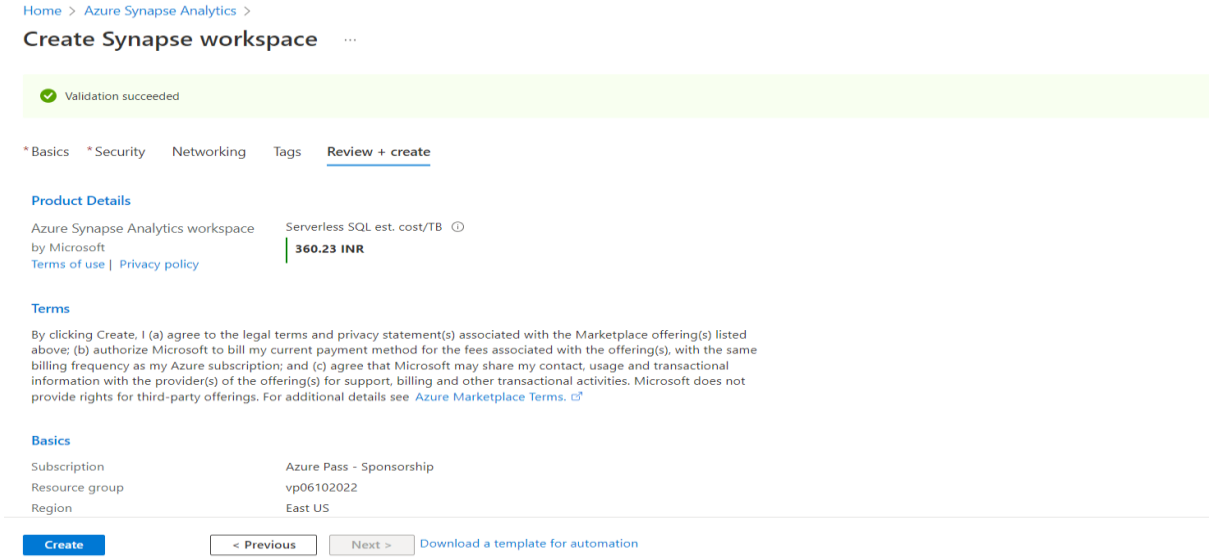


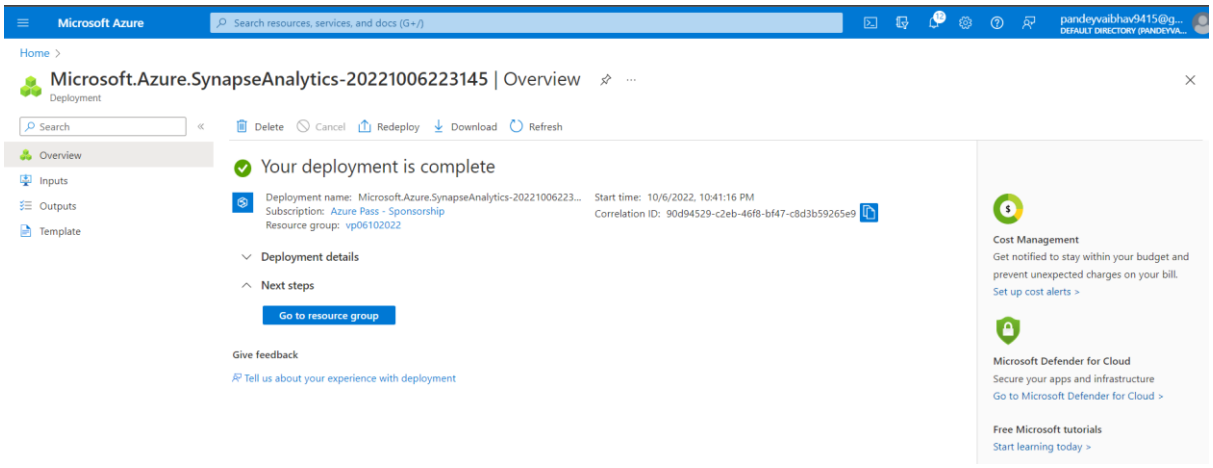

Home > Azure Synapse Analytics >
Create Synapse workspace
*Basics *Security Networking Tags Review + create
Configure security options for your workspace.
Authentication
Choose the authentication method for access to workspace resources such as SQL pools. The authentication method can be changed later on. Learn more
Authentication method
Use both local and Azure Active Directory (Azure AD) authentication
Use only Azure Active Directory (Azure AD) authentication
SQL Server admin login *
sqladminuser
SQL Password
Confirm password
System assigned managed identity permission
Select to grant the workspace network access to the Data Lake Storage Gen2 account using the workspace system identity. Learn more
Allow network access to Data Lake Storage Gen2 account.
The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a
Review + create
< Previous
Next: Networking >

Default

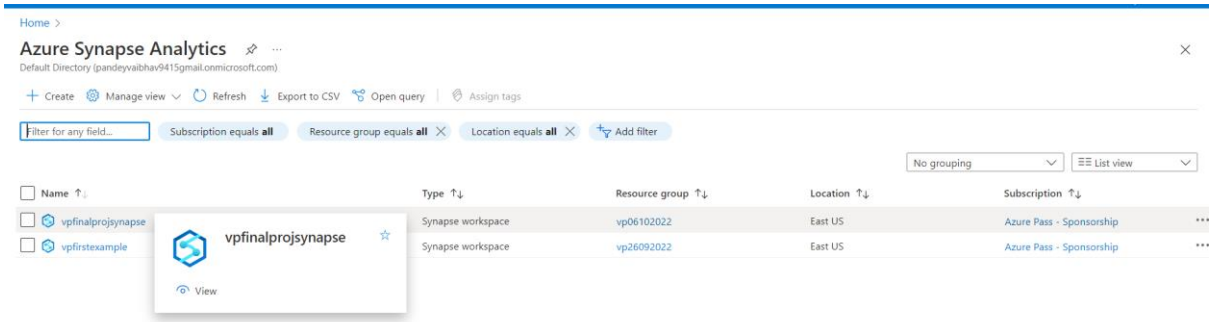

Home > Azure Synapse Analytics >
Create Synapse workspace
*Basics *Security Networking Tags Review + create
Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. learn more
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
:
Synapse workspace

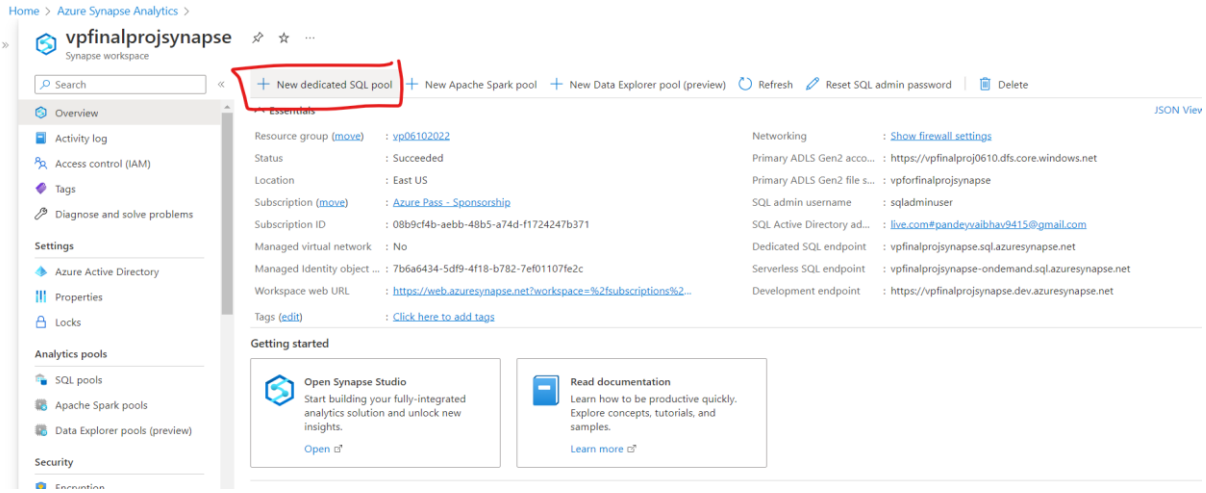

Home > Azure Synapse Analytics >
Create Synapse workspace
*Basics *Security Networking Tags Review + create
Configure networking options for your workspace.
Managed virtual network
Choose whether to set up a dedicated Azure Synapse-managed virtual network for your workspace. Learn more
Managed virtual network
Enable
Disable
To control public network access to your Synapse workspace, you must enable managed virtual network.
Firewall rules
Azure Synapse Studio and other client tools will only connect to the workspace endpoints if this setting is selected. Connections from specific IP addresses or all Azure services can be allowed or disallowed after the workspace is provisioned.
Allow connections from all IP addresses to your workspace's endpoints. You can restrict these permissions to just Azure datacenter IP addresses and/or specific IP address ranges after creating the workspace.
Allow connections from all IP addresses
Encrypted connections
Review + create
< Previous
Next: Tags >

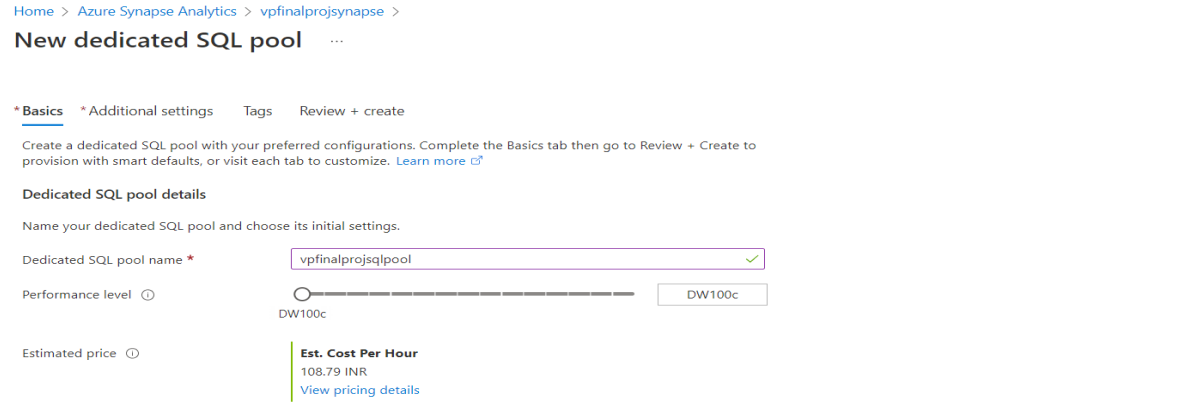


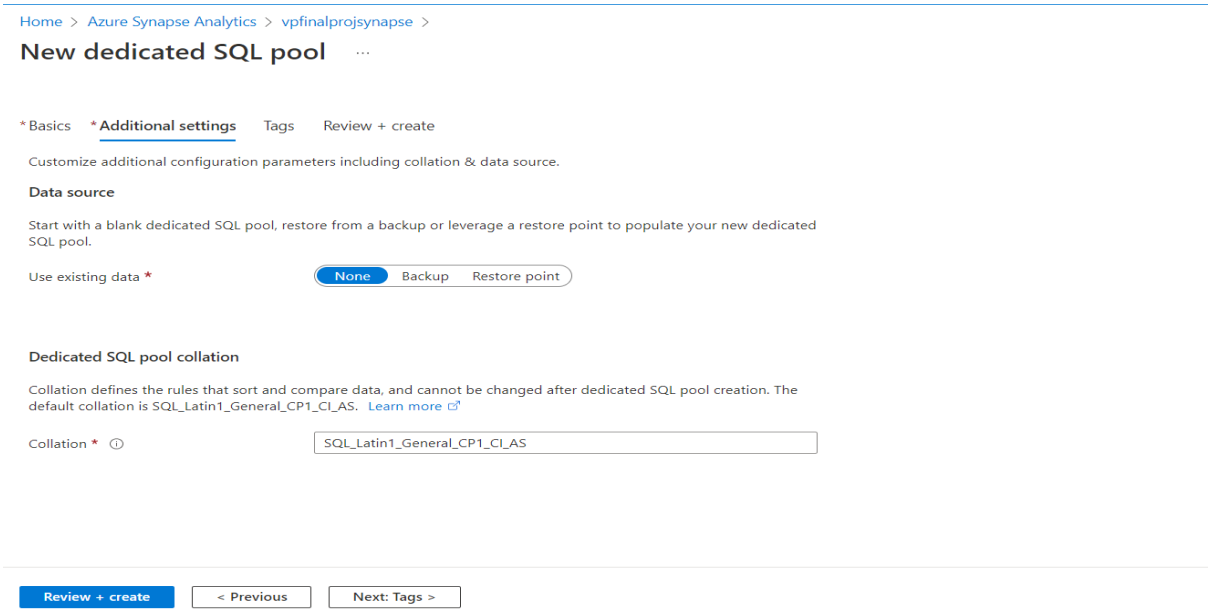


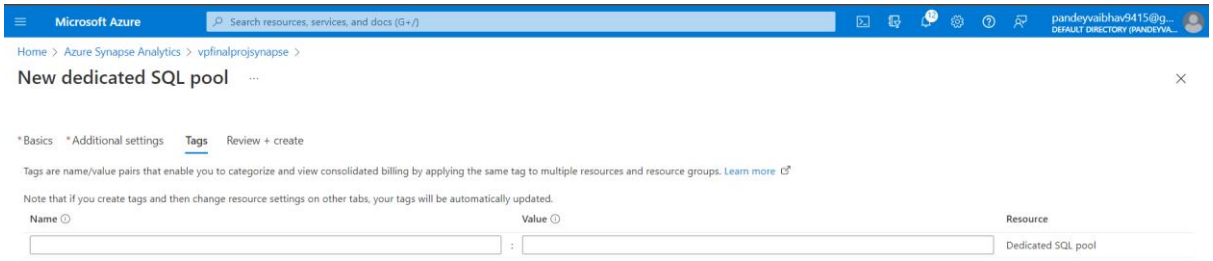
For making a dedicated SQL Pool

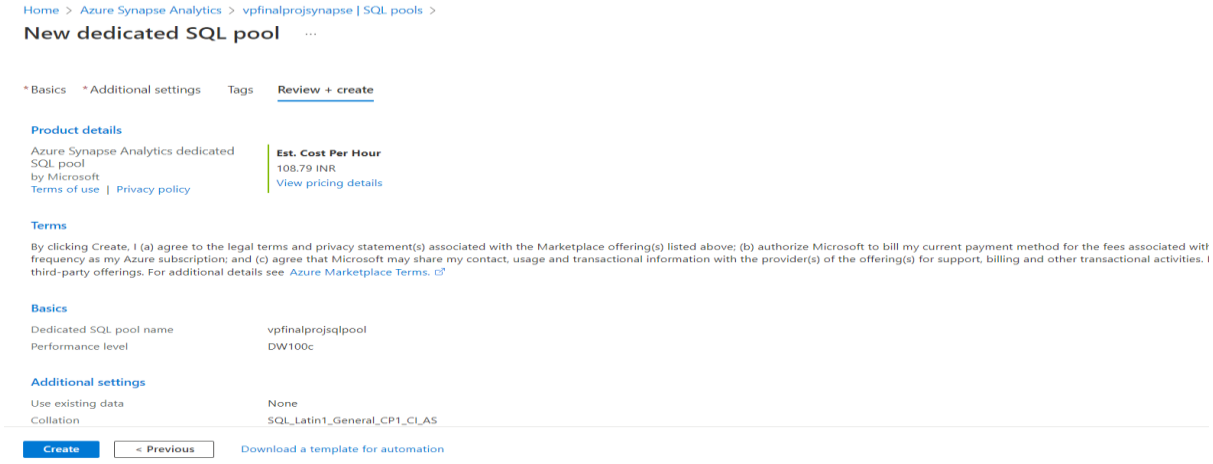


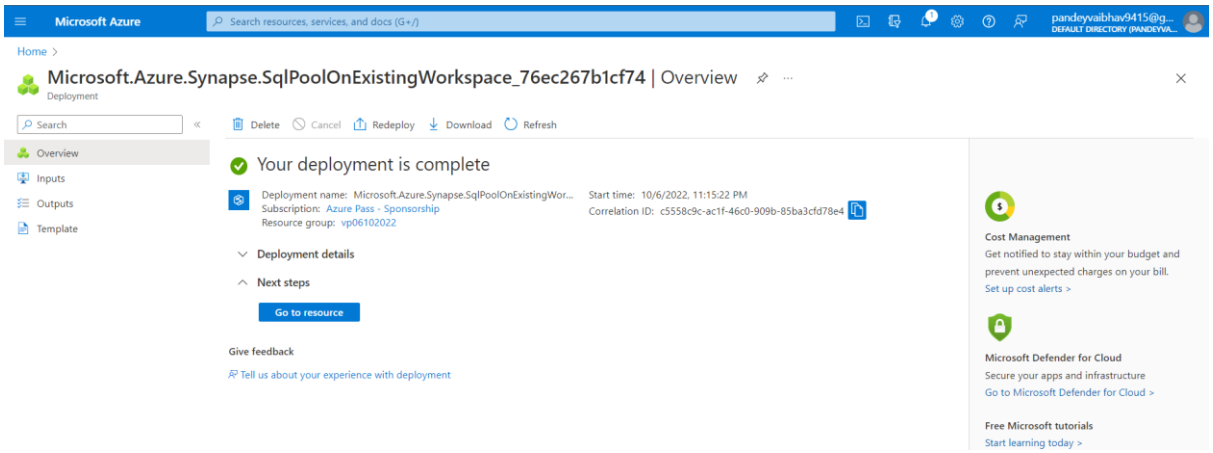

Home > Azure Synapse Analytics >
vpfinalprojsynapse
Synapse workspace
Search
New dedicated SQL pool
New Apache Spark pool
New Data Explorer pool (preview)
Refresh
Reset SQL admin password
Delete
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Azure Active Directory
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Data Explorer pools (preview)
Security
Essentials
JSON View
Resource group (move) : vp06102022
Status : Succeeded
Location : East US
Subscription (move) : Azure Pass - Sponsorship
Subscription ID : 08b9cf4b-aebb-48b5-a74d-f1724247b371
Managed virtual network : No
Managed Identity object ... : 7b6a6434-5df9-4f18-b782-7ef01107fe2c
Workspace web URL : https://web.azuresynapse.net?workspace=%2fsubscriptions%2...
Tags (edit) : Click here to add tags
Networking : Show firewall settings
Primary ADLS Gen2 acco... : https://vpfinalproj0610.dfs.core.windows.net
Primary ADLS Gen2 file s... : vpforfinalprojsynapse
SQL admin username : sqladminuser
SQL Active Directory ad... : live.com#pandeyvaibhav9415@gmail.com
Dedicated SQL endpoint : vpfinalprojsynapse.sql.azuresynapse.net
Serverless SQL endpoint : vpfinalprojsynapse-ondemand.sql.azuresynapse.net
Development endpoint : https://vpfinalprojsynapse.dev.azuresynapse.net
Getting started
Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more

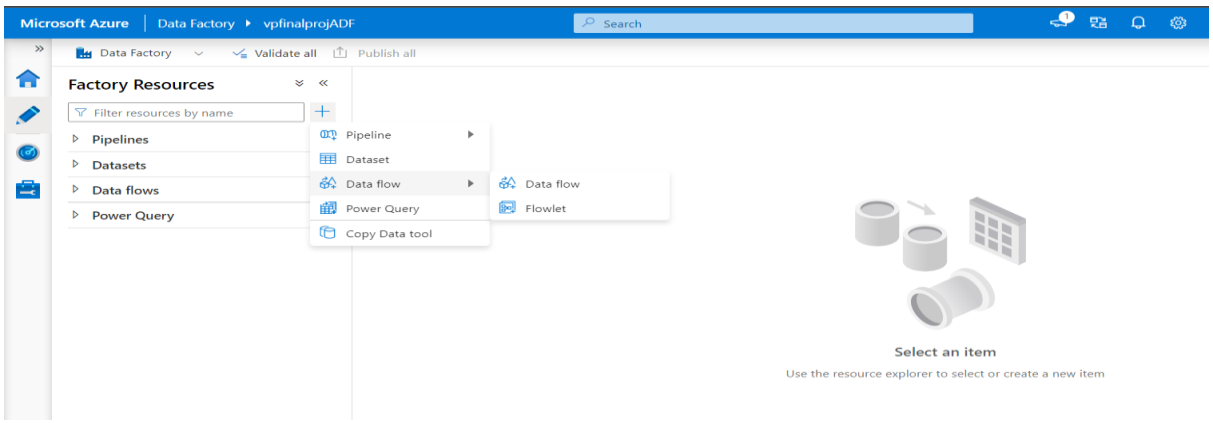

Home > Azure Synapse Analytics > vpfinalprojsynapse >
New dedicated SQL pool
*Basics
*Additional settings
Tags
Review + create
Create a dedicated SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. Learn more
Dedicated SQL pool details
Name your dedicated SQL pool and choose its initial settings.
Dedicated SQL pool name *
vpfinalprojsqlpool
Performance level
DW100c
DW100c
Estimated price
Est. Cost Per Hour
108.79 INR
View pricing details
Review + create
Next: Additional settings >

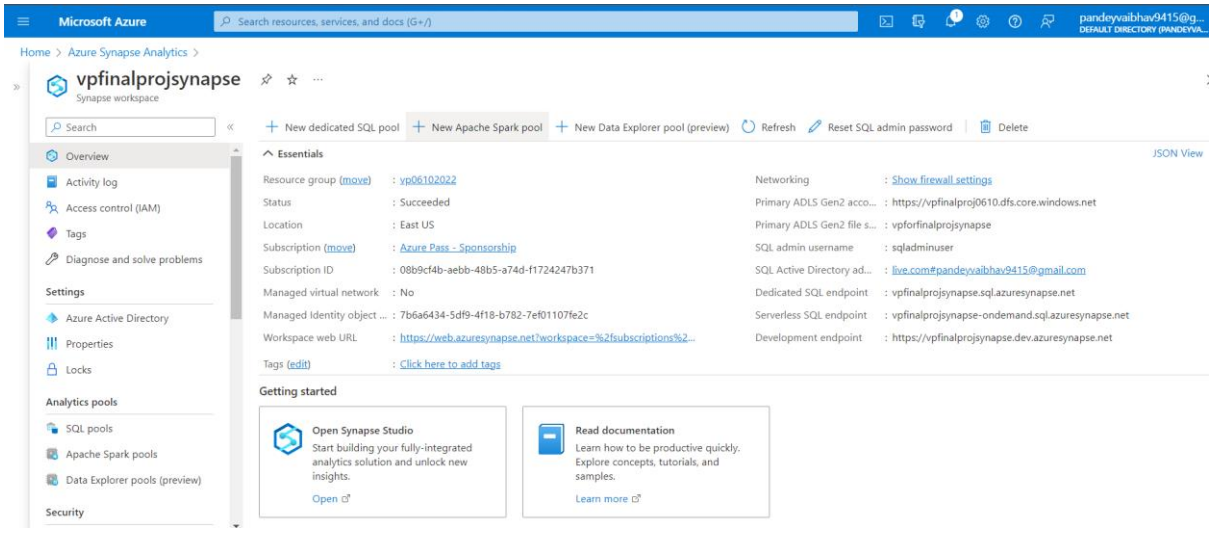

Home > Azure Synapse Analytics > vpfinalprojsynapse >
New dedicated SQL pool
*Basics
*Additional settings
Tags
Review + create
Customize additional configuration parameters including collation & data source.
Data source
Start with a blank dedicated SQL pool, restore from a backup or leverage a restore point to populate your new dedicated SQL pool.
Use existing data *
None
Backup
Restore point
Dedicated SQL pool collation
Collation defines the rules that sort and compare data, and cannot be changed after dedicated SQL pool creation. The default collation is SQL_Latin1_General_CP1_CI_AS. Learn more
Collation *
SQL_Latin1_General_CP1_CI_AS
Review + create
< Previous
Next: Tags >

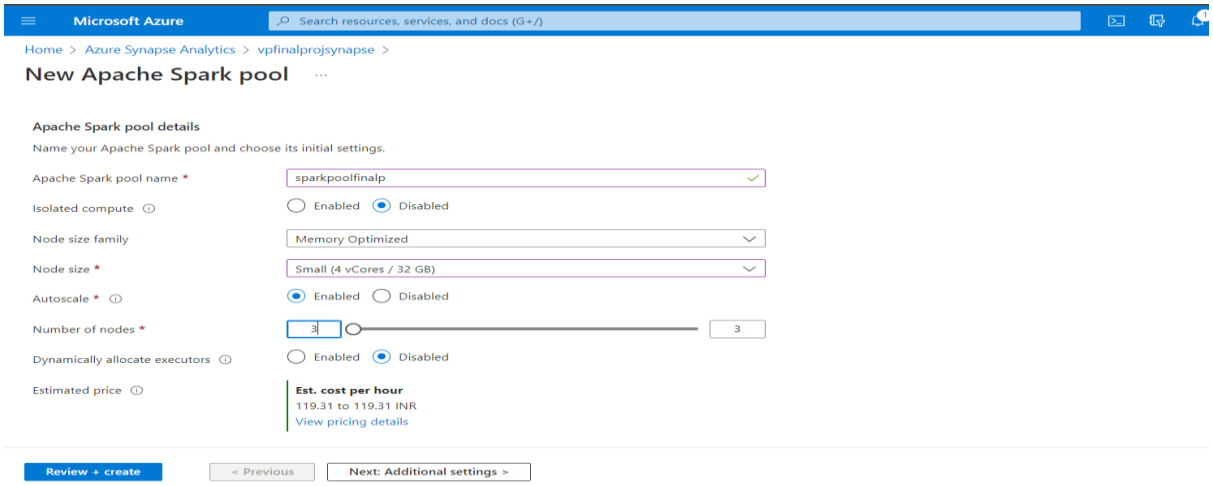

Microsoft Azure
Search resources, services, and docs (G+/)
pandeyvaibhav9415@g...
DEFAULT DIRECTORY (PANDEYVA...
Home > Azure Synapse Analytics > vpfinalprojsynapse >
New dedicated SQL pool
*Basics *Additional settings Tags Review + create
Tags are name/value pairs that enable you to categorize and view consolidated billing by applying the same tag to multiple resources and resource groups. Learn more
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
Dedicated SQL pool

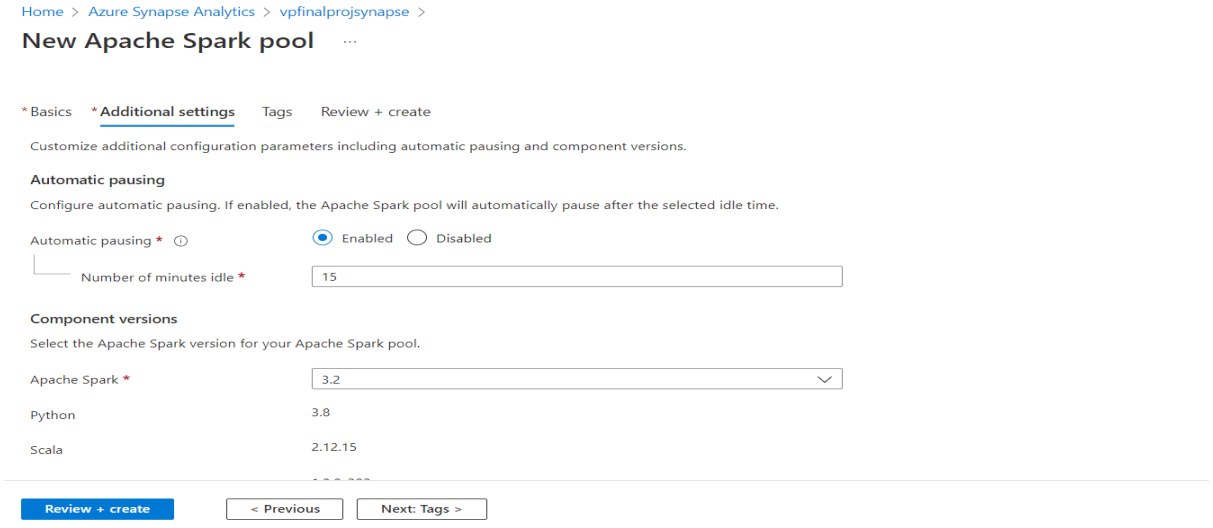

Home > Azure Synapse Analytics > vpfinalprojsynapse | SQL pools >
New dedicated SQL pool
*Basics *Additional settings Tags Review + create
Product details
Azure Synapse Analytics dedicated SQL pool
by Microsoft
Terms of use | Privacy policy
Est. Cost Per Hour
108.79 INR
View pricing details
Terms
By 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
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.
third-party offerings. For additional details see Azure Marketplace Terms.
Basics
Dedicated SQL pool name vpfinalprojsqlpool
Performance level DW100c
Additional settings
Use existing data None
Collation SQL_Latin1_General_CP1_CI_AS
Create
< Previous
Download a template for automation

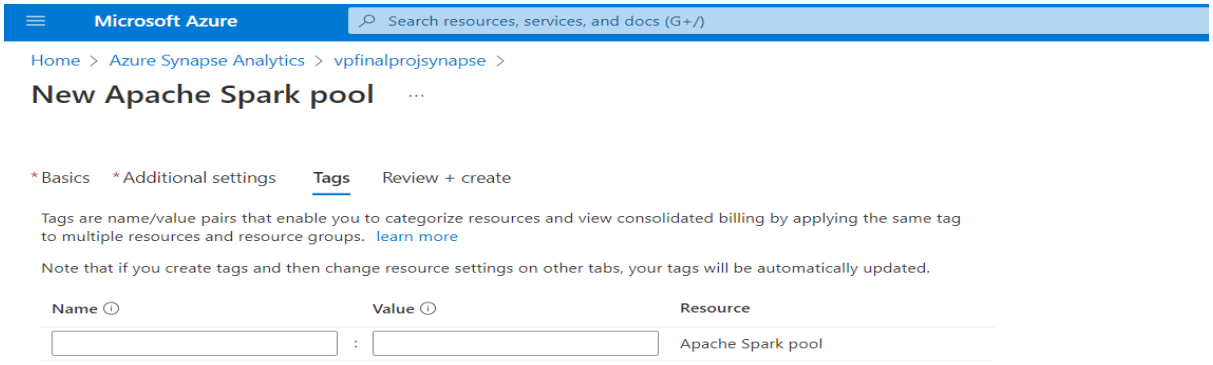

Microsoft Azure
Search resources, services, and docs (G+/)
pandeyvaibhav9415@g...
DEFAULT DIRECTORY (PANDEYVA...
Home >
Microsoft.Azure.Synapse.SqlPoolOnExistingWorkspace_76ec267b1cf74 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete Cancel Redeploy Download Refresh
Your deployment is complete
Deployment name: Microsoft.Azure.Synapse.SqlPoolOnExistingWor...
Subscription: Azure Pass - Sponsorship
Resource group: vp06102022
Start time: 10/6/2022, 11:15:22 PM
Correlation ID: c5558c9c-ac1f-46c0-909b-85ba3cfd78e4
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 >
Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >
Free Microsoft tutorials
Start learning today >

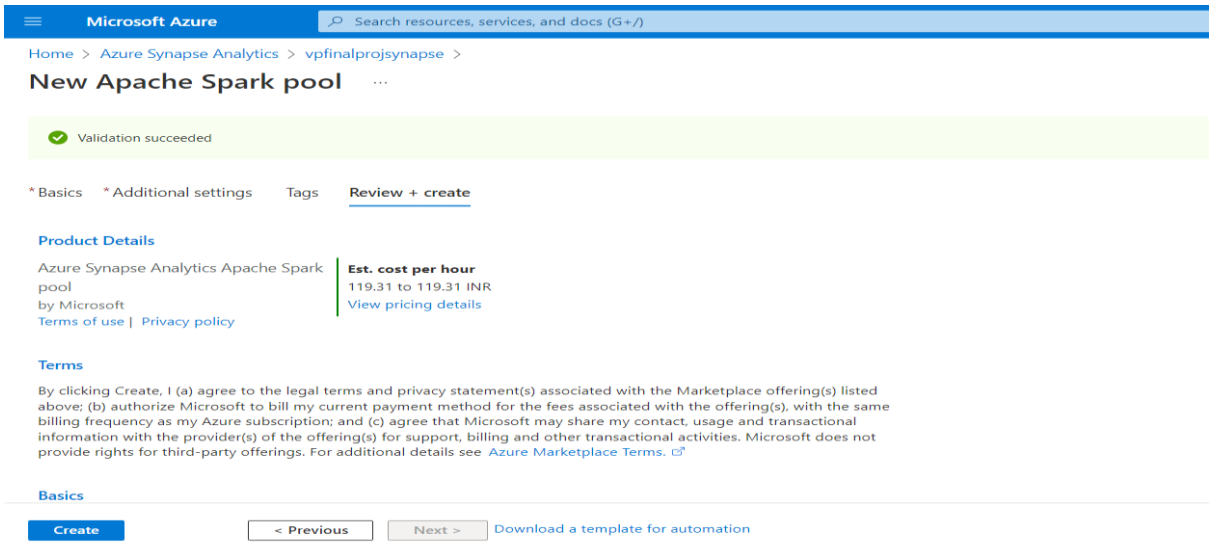

Microsoft Azure | Data Factory ▸ vpfinalprojADF
Search
Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
Datasets
Data flows
Power Query
Pipeline
Dataset
Data flow
Power Query
Copy Data tool
Data flow
Flowlet
Select an item
Use the resource explorer to select or create a new item

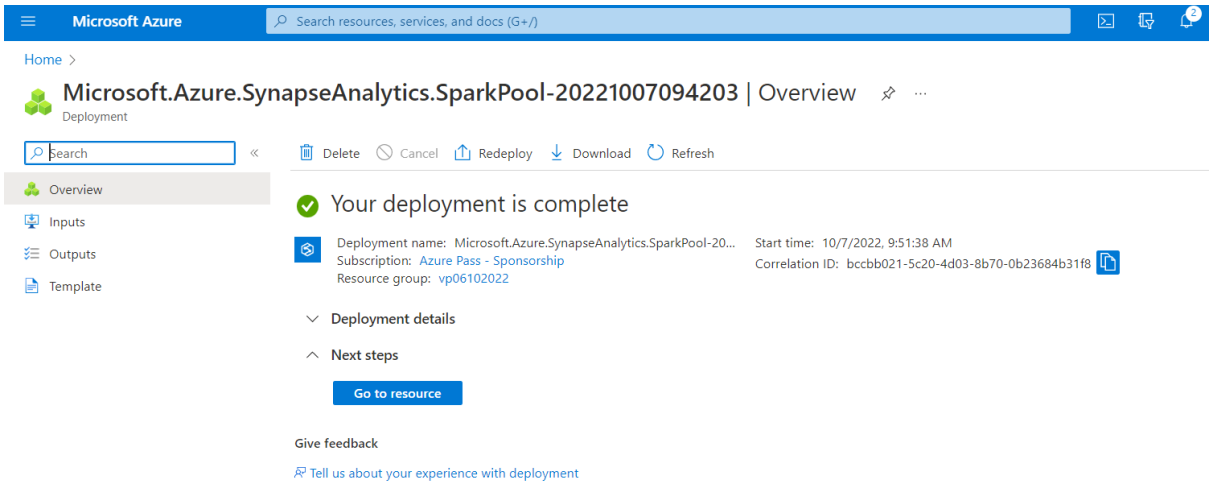

Microsoft Azure
Search resources, services, and docs (G+/)
pandeyvaibhav9415@g...
DEFAULT DIRECTORY (PANDEYVA...
Home > Azure Synapse Analytics >
vpfinalprojsynapse
Synapse workspace
Search
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Settings
Azure Active Directory
Properties
Locks
Analytics pools
SQL pools
Apache Spark pools
Data Explorer pools (preview)
Security
New dedicated SQL pool
New Apache Spark pool
New Data Explorer pool (preview)
Refresh
Reset SQL admin password
Delete
Essentials
JSON View
Resource group (move) : vp06102022
Status : Succeeded
Location : East US
Subscription (move) : Azure Pass - Sponsorship
Subscription ID : 08b9cf4b-aebb-48b5-a74d-f1724247b371
Managed virtual network : No
Managed Identity object ... : 7b6a6434-5df9-4f18-b782-7ef01107fe2c
Workspace web URL : https://web.azuresynapse.net?workspace=%2fsubscriptions%2...
Tags (edit) : Click here to add tags
Networking : Show firewall settings
Primary ADLS Gen2 acco... : https://vpfinalproj0610.dfs.core.windows.net
Primary ADLS Gen2 file s... : vpforfinalprojsynapse
SQL admin username : sqladminuser
SQL Active Directory ad... : live.com#pandeyvaibhav9415@gmail.com
Dedicated SQL endpoint : vpfinalprojsynapse.sql.azuresynapse.net
Serverless SQL endpoint : vpfinalprojsynapse-ondemand.sql.azuresynapse.net
Development endpoint : https://vpfinalprojsynapse.dev.azuresynapse.net
Getting started
Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more


Microsoft Azure
Search resources, services, and docs (G+/)
Home > Azure Synapse Analytics > vpfinalprojsynapse >
New Apache Spark pool
Apache Spark pool details
Name your Apache Spark pool and choose its initial settings.
Apache Spark pool name * sparkpoolfinalp
Isolated compute Enabled Disabled
Node size family Memory Optimized
Node size * Small (4 vCores / 32 GB)
Autoscale * Enabled Disabled
Number of nodes * 3 3
Dynamically allocate executors Enabled Disabled
Estimated price
Est. cost per hour
119.31 to 119.31 INR
View pricing details
Review + create
< Previous
Next: Additional settings >


Home > Azure Synapse Analytics > vpfinalprojsynapse >
New Apache Spark pool
* Basics * Additional settings Tags Review + create
Customize additional configuration parameters including automatic pausing and component versions.
Automatic pausing
Configure automatic pausing. If enabled, the Apache Spark pool will automatically pause after the selected idle time.
Automatic pausing * Enabled Disabled
Number of minutes idle * 15
Component versions
Select the Apache Spark version for your Apache Spark pool.
Apache Spark * 3.2
Python 3.8
Scala 2.12.15
Review + create
< Previous
Next: Tags >


Microsoft Azure
Search resources, services, and docs (G+/)
Home > Azure Synapse Analytics > vpfinalprojsynapse >
New Apache Spark pool
* Basics * Additional settings Tags Review + create
Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. learn more
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
Apache Spark pool


Microsoft Azure
Search resources, services, and docs (G+/)
Home > Azure Synapse Analytics > vpfinalprojsynapse >
New Apache Spark pool
Validation succeeded
* Basics * Additional settings Tags Review + create
Product Details
Azure Synapse Analytics Apache Spark pool
by Microsoft
Terms of use | Privacy policy
Est. cost per hour
119.31 to 119.31 INR
View pricing details
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
Create
< Previous
Next >
Download a template for automation


Microsoft Azure
Search resources, services, and docs (G+/)
Home >
Microsoft.Azure.SynapseAnalytics.SparkPool-20221007094203 | Overview
Deployment
Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh
Your deployment is complete
Deployment name: Microsoft.Azure.SynapseAnalytics.SparkPool-20...
Subscription: Azure Pass - Sponsorship
Resource group: vp06102022
Start time: 10/7/2022, 9:51:38 AM
Correlation ID: bccbb021-5c20-4d03-8b70-0b23684b31f8
Deployment details
Next steps
Go to resource
Give feedback
Tell us about your experience with deployment

● Use the different Azure data factory tools to build a pipeline (SQL Database-> Copy-> ADLS Gen 2 -> Transform using Databricks -> Copy to Synapse DW).

● Use Databricks notebook for mounting ADLS Gen 2 storage, transforming the data (clean, join, filter, aggregate, pivot) and persist result to ADLS.

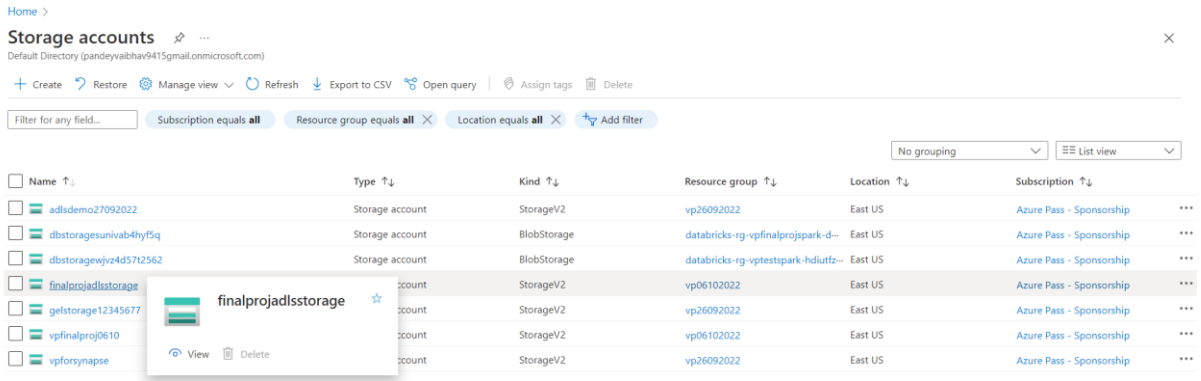
● Schedule and Monitor the pipeline and activity runs.

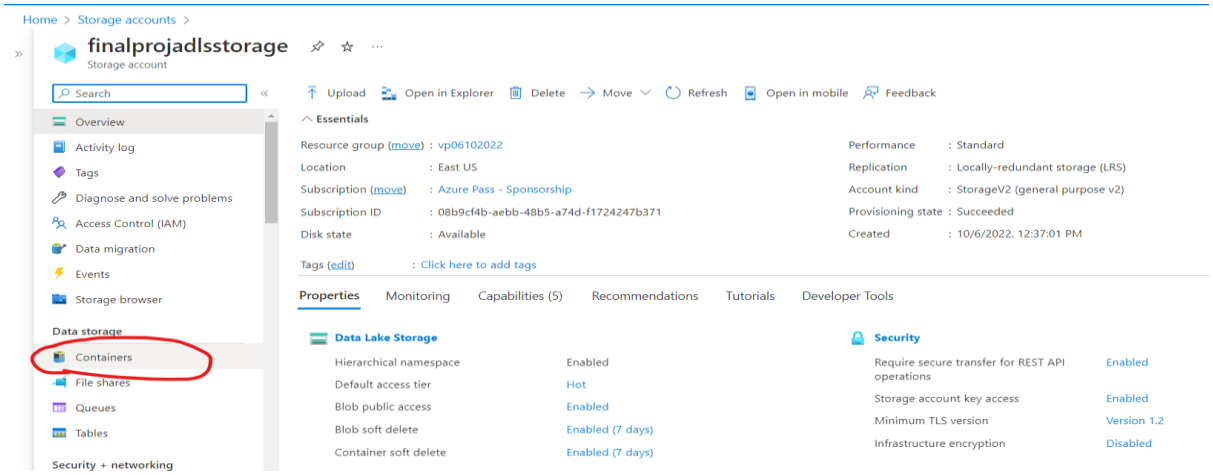
Questions that need to be answered/Evaluation steps while building the ETL Pipeline :

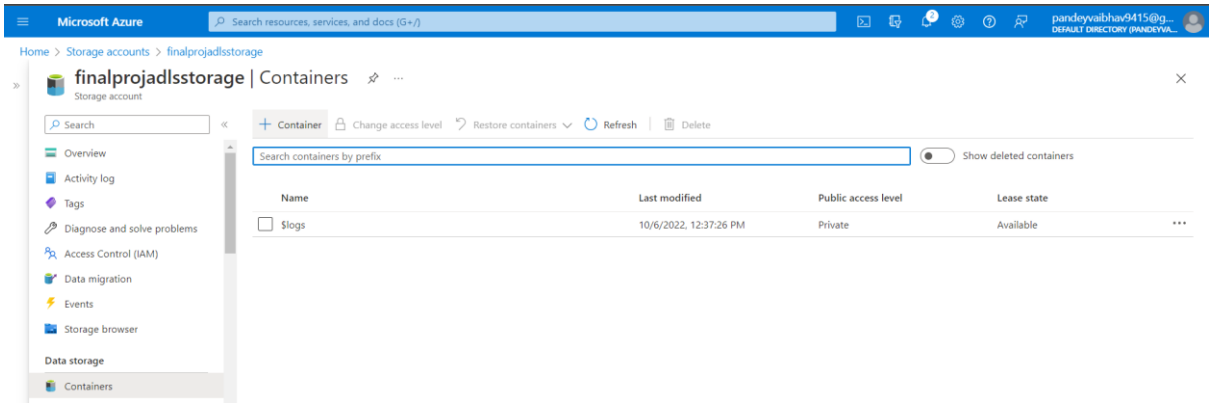
Task 1: Create a dataflow with the following requirement:

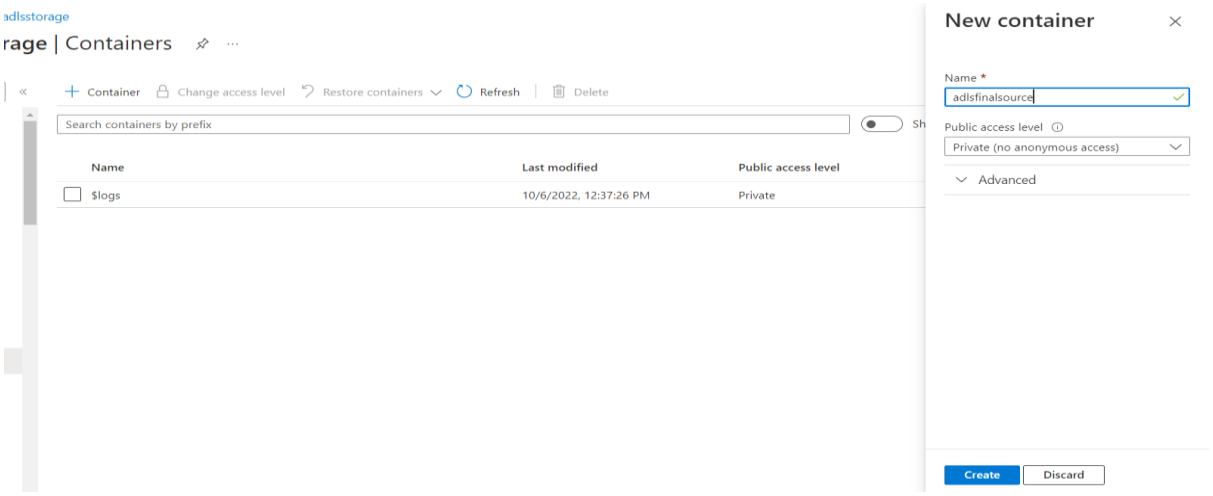
1. Create a data stream named CleaningGenreRomance and perform data cleansing on the Genre column using Derived Column and case expression. (While collecting data it was observed that some genres have spelling mistakes like romance, Romence for Romance, comedy, Comdy for Comedy.)

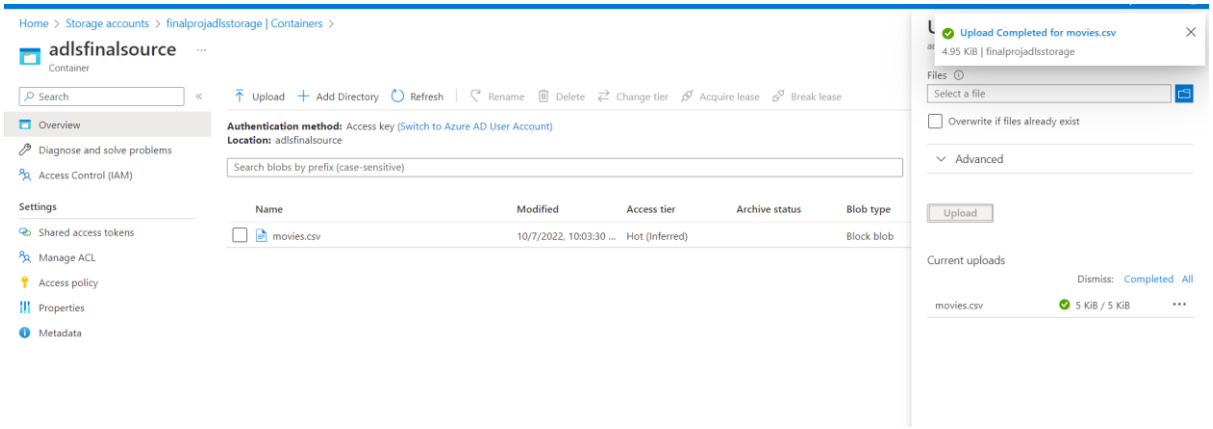
Step 1 : First store movies data in ADLS storage account.



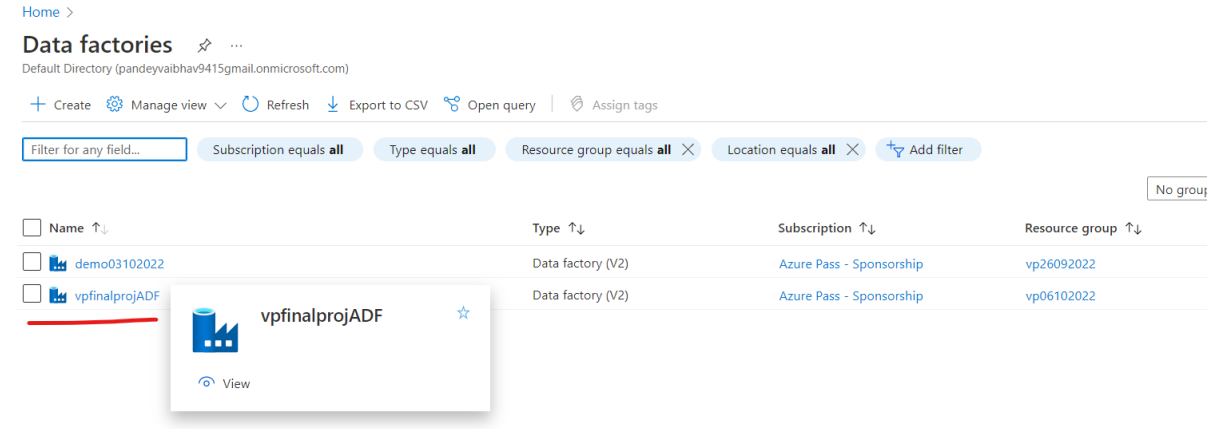

Home > Storage accounts >
finalprojadlsstorage
Storage account
Search
Upload
Open in Explorer
Delete
Move
Refresh
Open in mobile
Feedback
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Data storage
Containers
File shares
Queues
Tables
Security + networking
Essentials
Resource group (move) : vp06102022
Location : East US
Subscription (move) : Azure Pass - Sponsorship
Subscription ID : 08b9cf4b-aebb-48b5-a74d-f1724247b371
Disk state : Available
Performance : Standard
Replication : Locally-redundant storage (LRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 10/6/2022, 12:37:01 PM
Tags (edit) : Click here to add tags
Properties
Monitoring
Capabilities (5)
Recommendations
Tutorials
Developer Tools
Data Lake Storage
Hierarchical namespace Enabled
Default access tier Hot
Blob public access Enabled
Blob soft delete Enabled (7 days)
Container soft delete Enabled (7 days)
Security
Require secure transfer for REST API operations Enabled
Storage account key access Enabled
Minimum TLS version Version 1.2
Infrastructure encryption Disabled

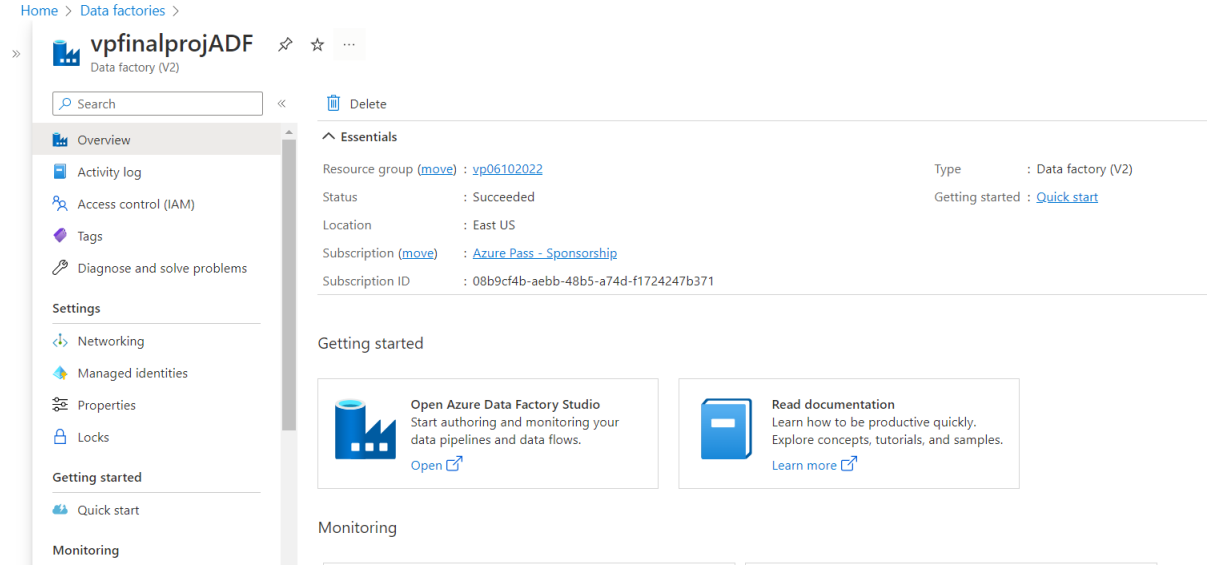

Microsoft Azure
Search resources, services, and docs (G+/)
Home > Storage accounts > finalprojadlsstorage
finalprojadlsstorage | Containers
Storage account
Container
Change access level
Restore containers
Refresh
Delete
Search containers by prefix
Show deleted containers
Name
Last modified
Public access level
Lease state
$logs
10/6/2022, 12:37:26 PM
Private
Available

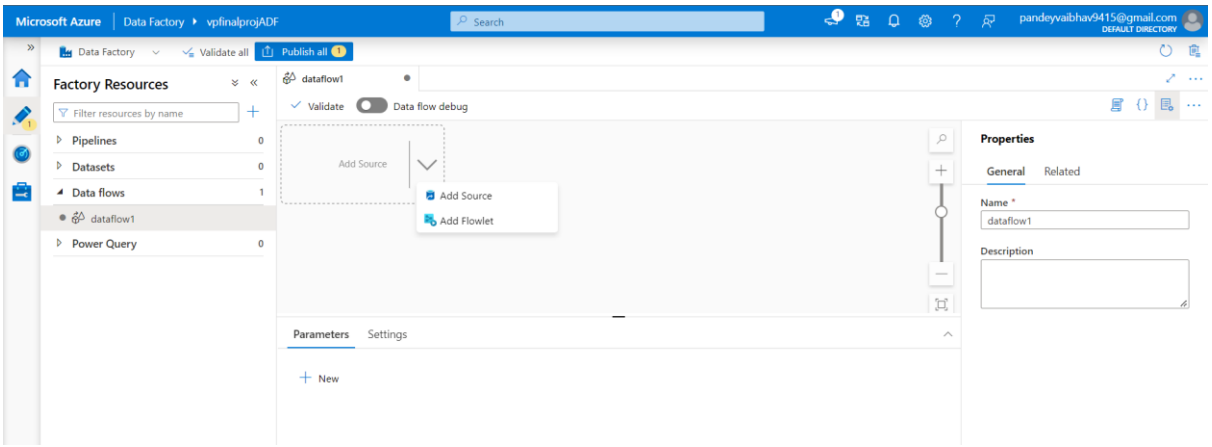

New container
Name *
adlsfinalsource
Public access level
Private (no anonymous access)
Advanced
Create
Discard
$logs
10/6/2022, 12:37:26 PM
Private

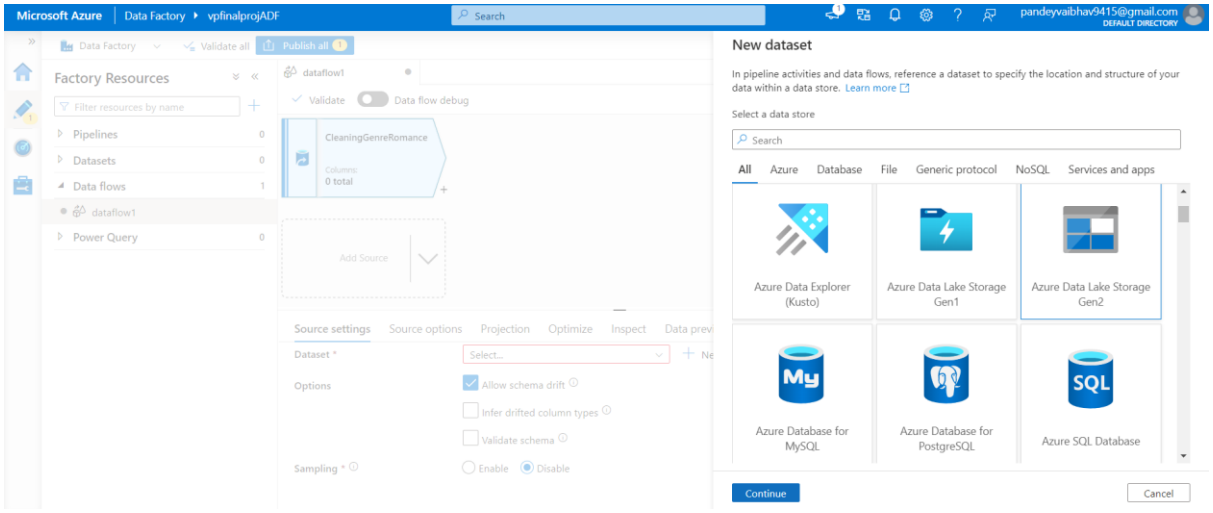

Home > Storage accounts > finalprojadlsstorage | Containers >
adlsfinalsource
Container
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Authentication method: Access key (Switch to Azure AD User Account)
Location: adlsfinalsource
Search blobs by prefix (case-sensitive)
Name
Modified
Access tier
Archive status
Blob type
movies.csv
10/7/2022, 10:03:30 ...
Hot (Inferred)
Block blob
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload Completed for movies.csv
4.95 KiB | finalprojadlsstorage
Files
Select a file
Overwrite if files already exist
Advanced
Upload
Current uploads
Dismiss: Completed All
movies.csv
5 KiB / 5 KiB

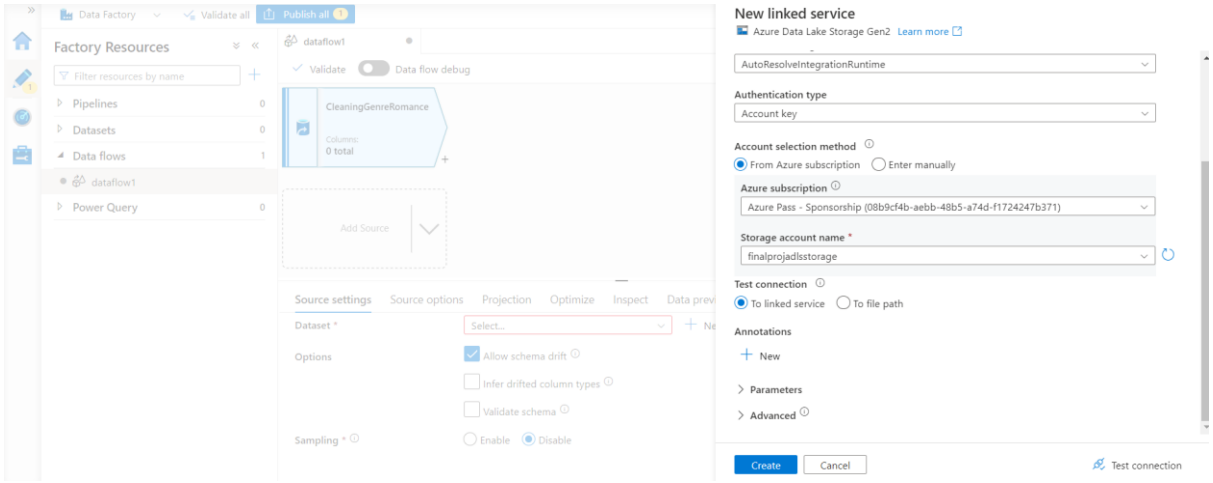
Now work on data flow

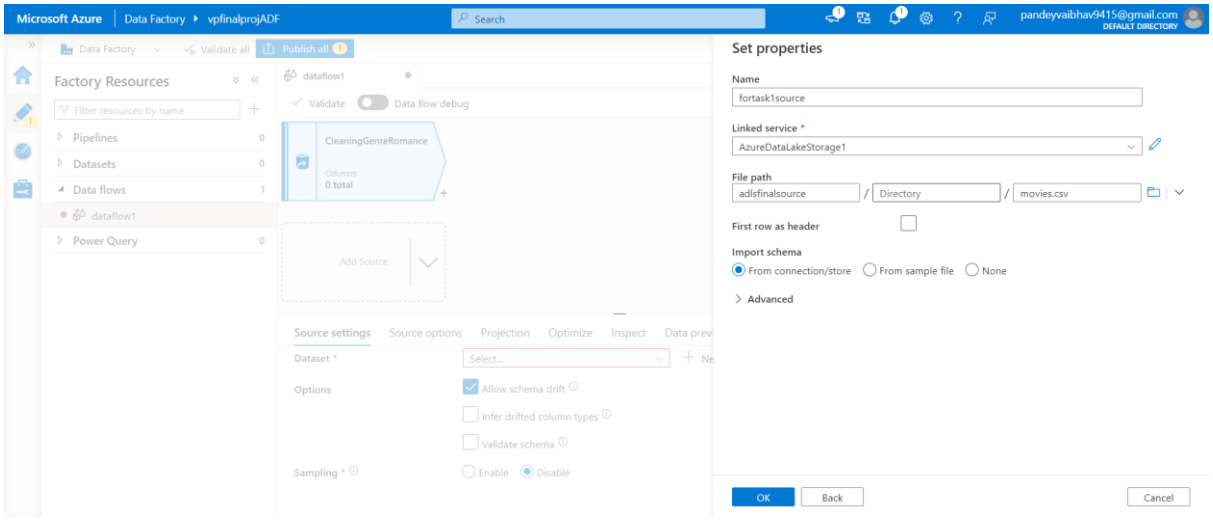


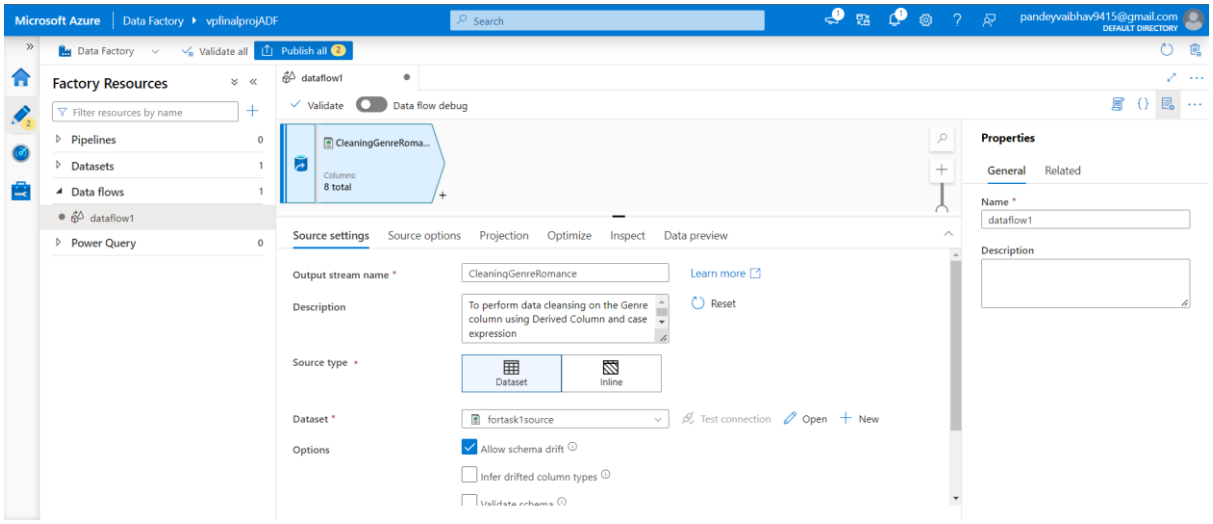




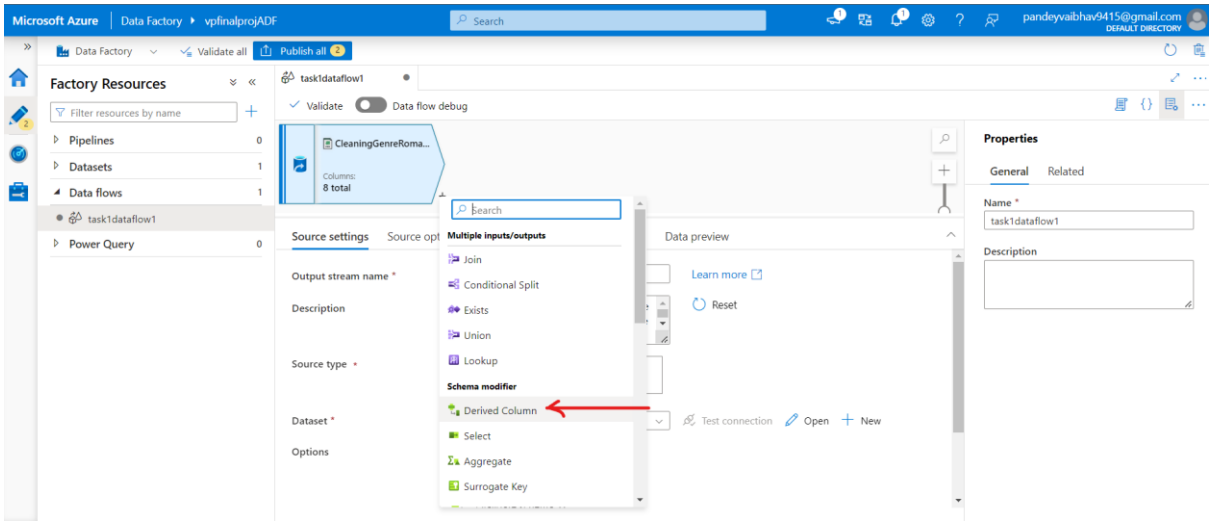

Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
Pipelines
Datasets
Data flows
dataflow1
Power Query
CleaningGenreRomance
Source settings
Source options
Projection
Optimize
Inspect
Dataset *
Options
Allow schema drift
Infer drifted column types
Validate schema
Sampling *
New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2
Azure Database for MySQL
Azure Database for PostgreSQL
Azure SQL Database
Continue
Cancel

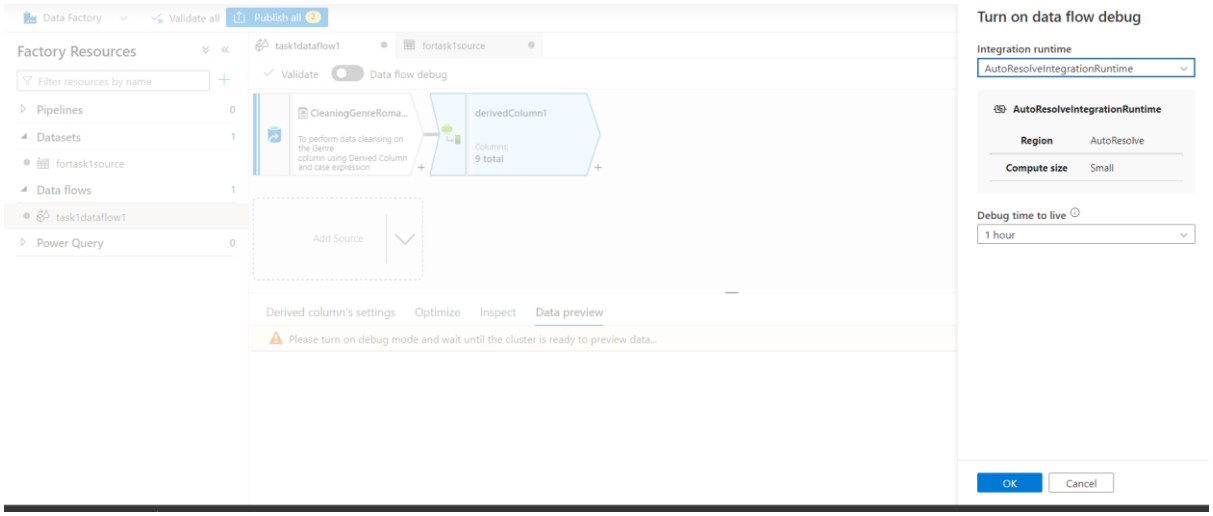

Factory Resources
dataflow1
CleaningGenreRomance
New linked service
Azure Data Lake Storage Gen2 Learn more
AutoResolveIntegrationRuntime
Authentication type
Account key
Account selection method
From Azure subscription
Enter manually
Azure subscription
Azure Pass - Sponsorship (08b9cf4b-aebb-48b5-a74d-f1724247b371)
Storage account name *
finalprojadlsstorage
Test connection
To linked service
To file path
Annotations
New
Parameters
Advanced
Create
Cancel
Test connection

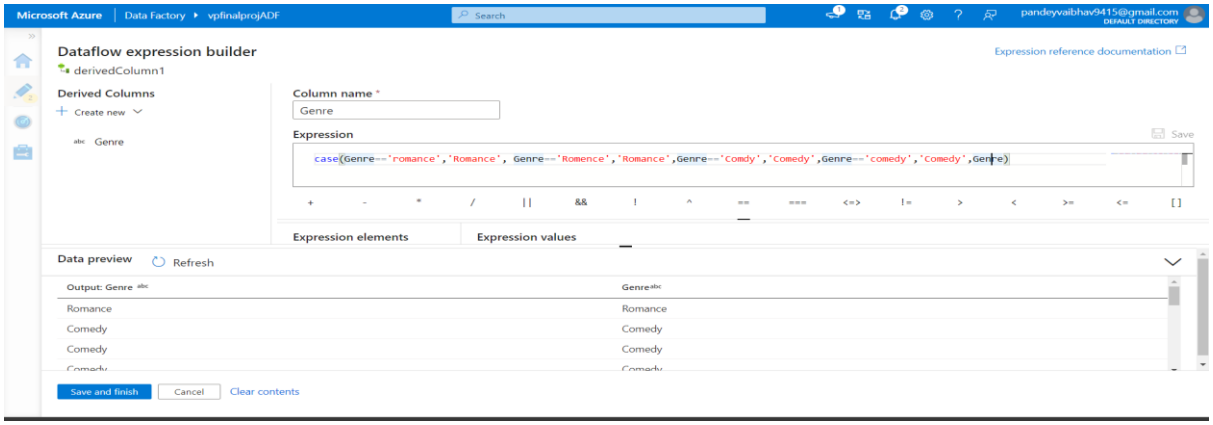

Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
dataflow1
CleaningGenreRomance
Set properties
Name
fortask1source
Linked service *
AzureDataLakeStorage1
File path
adlsfinalsource
Directory
movies.csv
First row as header
Import schema
From connection/store
From sample file
None
Advanced
OK
Back
Cancel

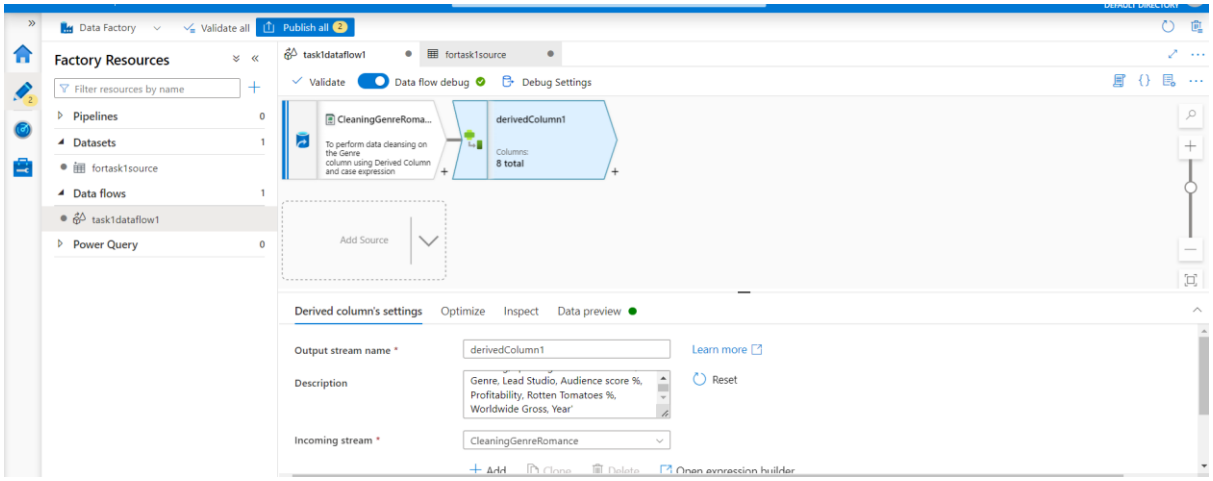


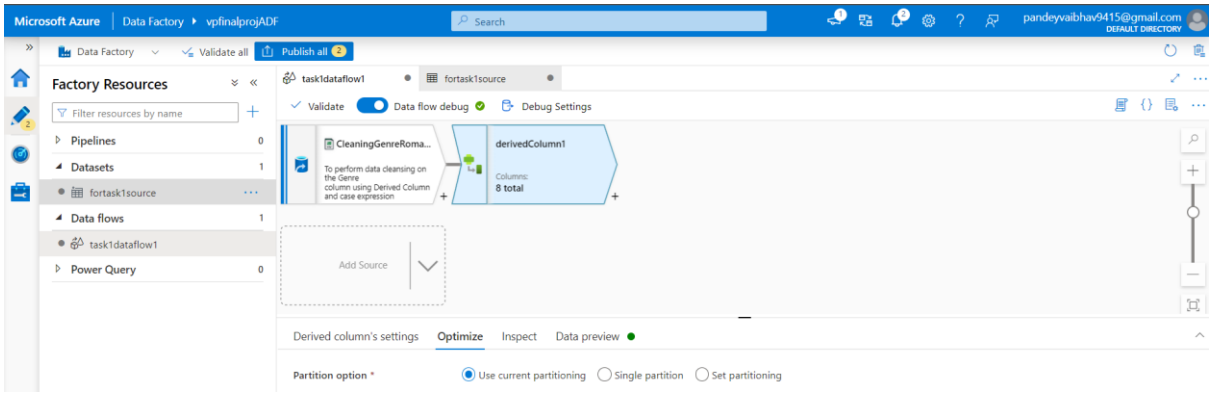
Now go for data cleansing:

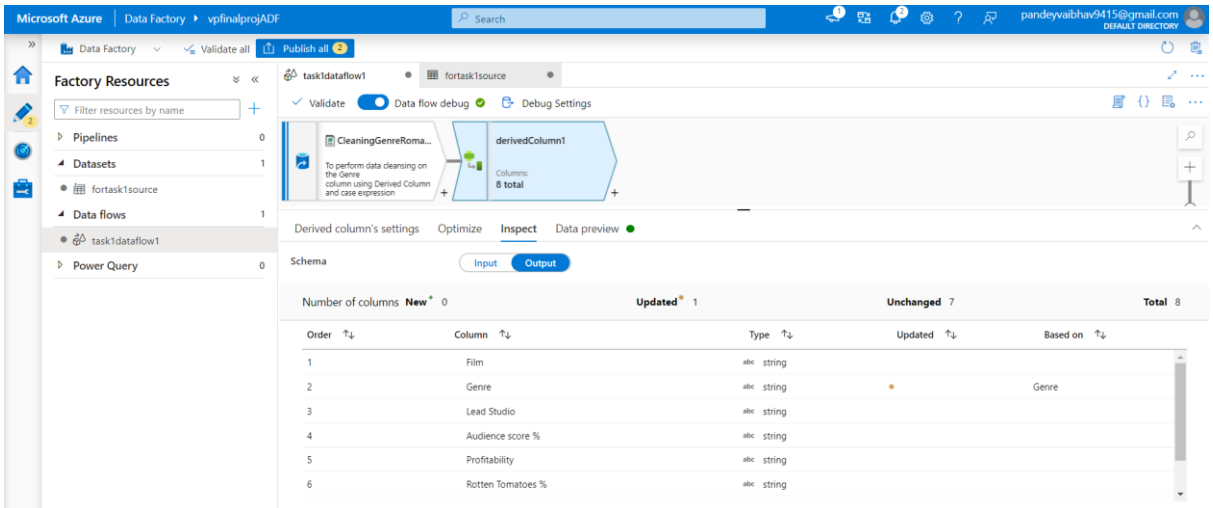


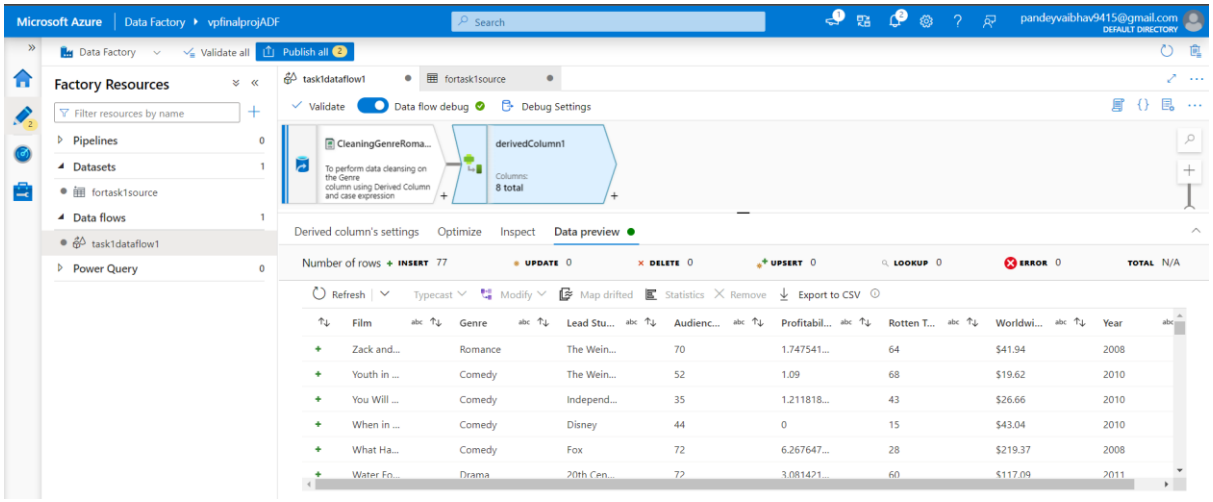


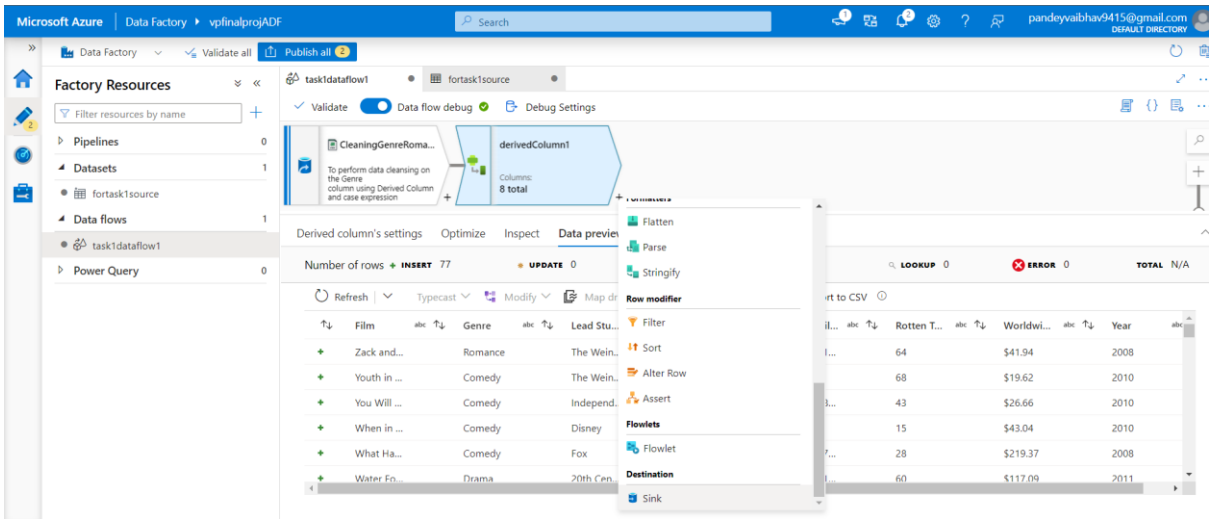

Microsoft Azure | Data Factory ▸ vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Dataflow expression builder
derivedColumn1
Expression reference documentation
Derived Columns
Create new
Genre
Column name *
Genre
Expression
Save
case(Genre=='romance','Romance', Genre=='Romence','Romance',Genre=='Comdy','Comedy',Genre=='comedy','Comedy',Genre)
+ - * / || && ! ^ == === <=> != > < >= <= []
Expression elements
Expression values
Data preview
Refresh
Output: Genre
Genre
Romance
Romance
Comedy
Comedy
Comedy
Comedy
Save and finish
Cancel
Clear contents

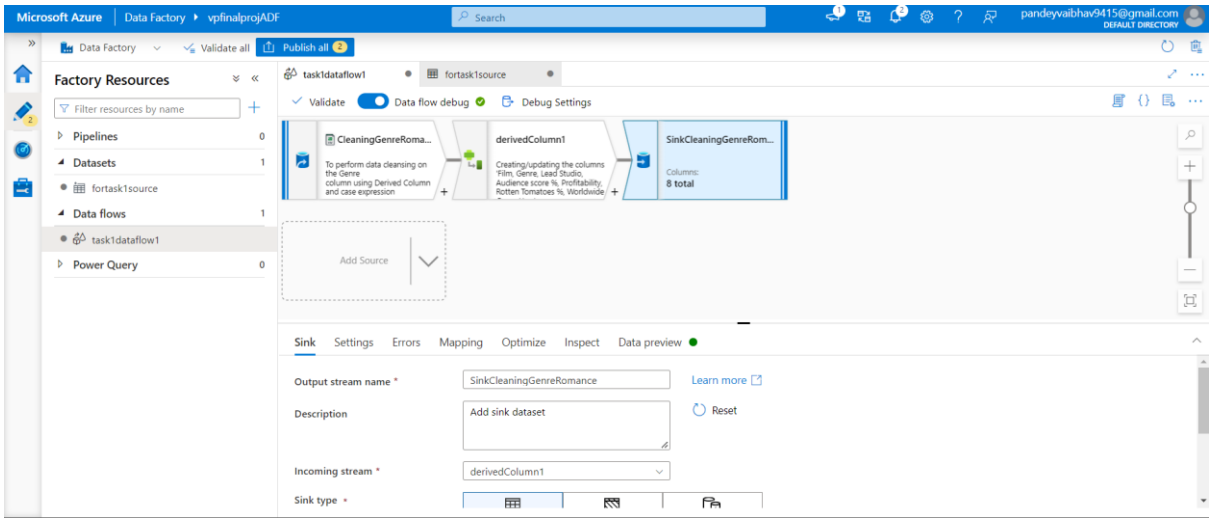

Data Factory
Validate all
Publish all 2
Factory Resources
Filter resources by name
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
To perform data cleansing on the Genre column using Derived Column and case expression
derivedColumn1
Columns:
8 total
Add Source
Derived column's settings
Optimize
Inspect
Data preview
Output stream name *
derivedColumn1
Learn more
Description
Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide Gross, Year'
Reset
Incoming stream *
CleaningGenreRomance
Add
Clone
Delete
Open expression builder

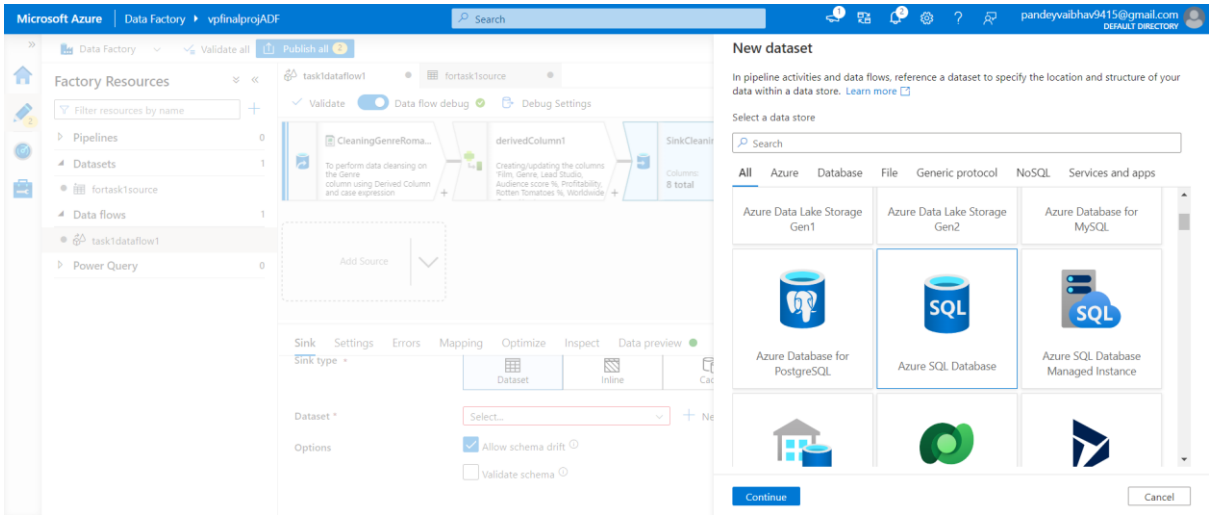

Microsoft Azure | Data Factory ▸ vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
Factory Resources
Filter resources by name
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
To perform data cleansing on the Genre column using Derived Column and case expression
derivedColumn1
Columns:
8 total
Add Source
Derived column's settings
Optimize
Inspect
Data preview
Partition option *
Use current partitioning
Single partition
Set partitioning

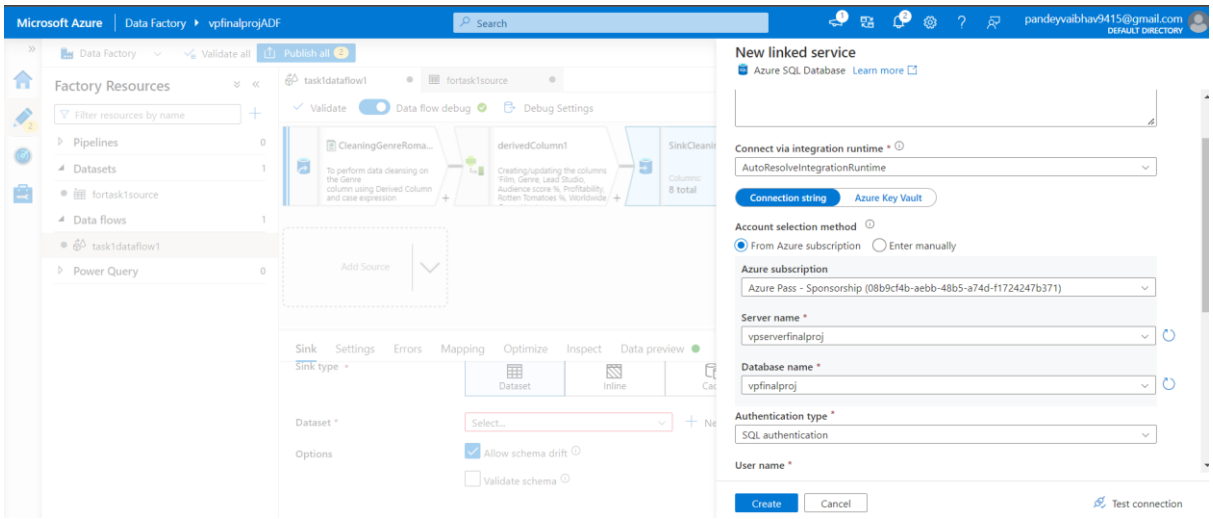

Microsoft Azure | Data Factory ▸ vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
Factory Resources
Filter resources by name
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
To perform data cleansing on the Genre column using Derived Column and case expression
derivedColumn1
Columns:
8 total
Derived column's settings
Optimize
Inspect
Data preview
Schema
Input
Output
Number of columns New 0
Updated 1
Unchanged 7
Total 8
Order
Column
Type
Updated
Based on
1 Film string
2 Genre string Genre
3 Lead Studio string
4 Audience score % string
5 Profitability string
6 Rotten Tomatoes % string

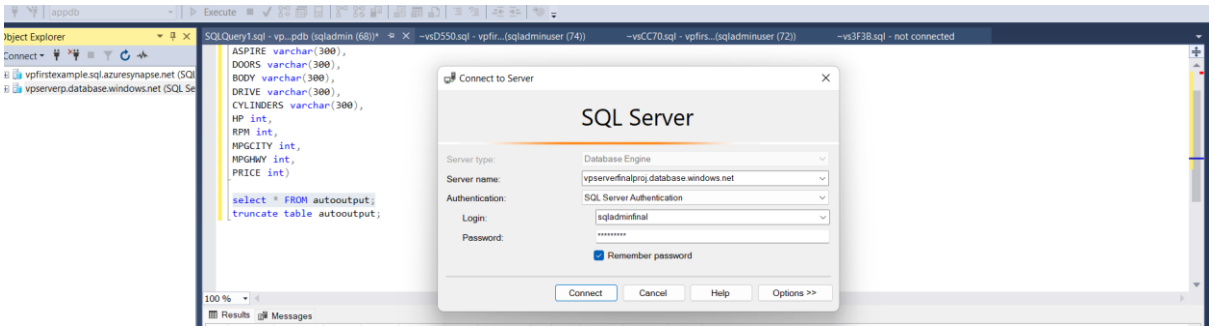

Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 2
Factory Resources
Filter resources by name
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
To perform data cleansing on the Genre column using Derived Column and case expression
derivedColumn1
Columns:
8 total
Derived column's settings
Optimize
Inspect
Data preview
Number of rows
INSERT 77
UPDATE 0
DELETE 0
UPSERT 0
LOOKUP 0
ERROR 0
TOTAL N/A
Refresh
Typecast
Modify
Map drifted
Statistics
Remove
Export to CSV
Film
Genre
Lead Stu...
Audienc...
Profitabil...
Rotten T...
Worldwi...
Year
Zack and... Romance The Wein... 70 1.747541... 64 $41.94 2008
Youth in ... Comedy The Wein... 52 1.09 68 $19.62 2010
You Will ... Comedy Independ... 35 1.211818... 43 $26.66 2010
When in ... Comedy Disney 44 0 15 $43.04 2010
What Ha... Comedy Fox 72 6.267647... 28 $219.37 2008
Water Fo... Drama 20th Cen... 72 3.081421... 60 $117.09 2011

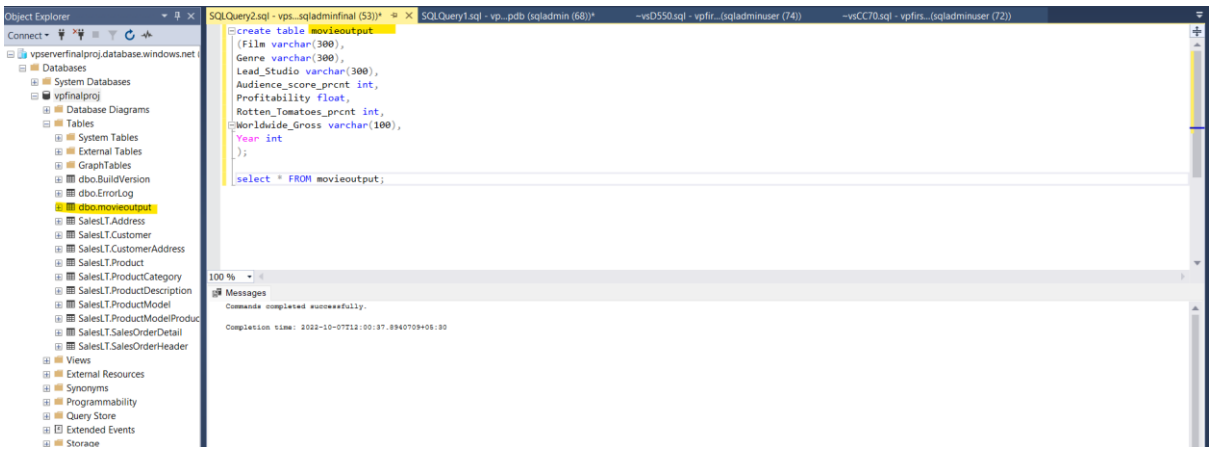

Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
CleaningGenreRoma...
derivedColumn1
Columns:
8 total
Flatten
Parse
Stringify
Row modifier
Filter
Sort
Alter Row
Assert
Flowlets
Flowlet
Destination
Sink


Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
Pipelines 0
Datasets 1
fortask1source
Data flows 1
task1dataflow1
Power Query 0
CleaningGenreRoma...
derivedColumn1
SinkCleaningGenreRom...
Columns:
8 total
Add Source
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
SinkCleaningGenreRomance
Learn more
Description
Add sink dataset
Reset
Incoming stream *
derivedColumn1
Sink type *


Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
Pipelines
Datasets
Data flows
Power Query
New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2
Azure Database for MySQL
Azure Database for PostgreSQL
Azure SQL Database
Azure SQL Database Managed Instance
Continue
Cancel


Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Factory Resources
New linked service
Azure SQL Database Learn more
Connect via integration runtime *
AutoResolveIntegrationRuntime
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Azure subscription
Azure Pass - Sponsorship (08b9cf4b-aebb-48b5-a74d-f1724247b371)
Server name *
vpserverfinalproj
Database name *
vpfinalproj
Authentication type *
SQL authentication
User name *
Create
Cancel
Test connection


Connect to Server
SQL Server
Server type:
Database Engine
Server name:
vpserverfinalproj.database.windows.net
Authentication:
SQL Server Authentication
Login:
sqladminfinal
Password:
Remember password
Connect
Cancel
Help
Options >>
ASPIRE varchar(300),
DOORS varchar(300),
BODY varchar(300),
DRIVE varchar(300),
CYLINDERS varchar(300),
HP int,
RPM int,
MPGCITY int,
MPGHWY int,
PRICE int)
select * FROM autooutput;
truncate table autooutput;

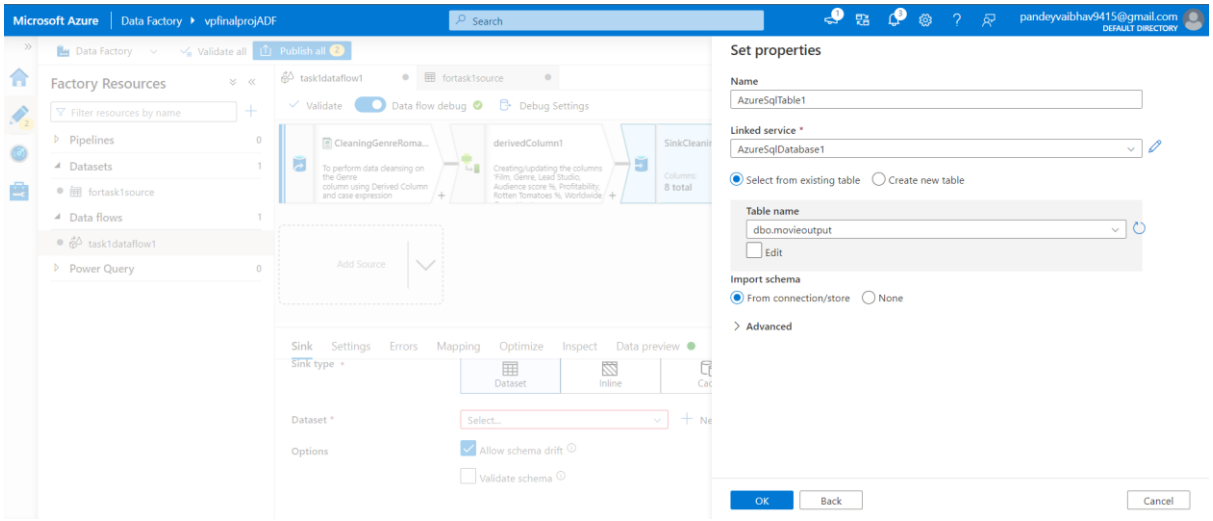

Object Explorer
vpserverfinalproj.database.windows.net
Databases
System Databases
vpfinalproj
Database Diagrams
Tables
dbo.movieoutput
create table movieoutput
(Film varchar(300),
Genre varchar(300),
Lead_Studio varchar(300),
Audience_score_prcnt int,
Profitability float,
Rotten_Tomatoes_prcnt int,
Worldwide_Gross varchar(100),
Year int
);
select * FROM movieoutput;
Messages
Commands completed successfully.
Completion time: 2022-10-07T12:00:37.8940709+05:30

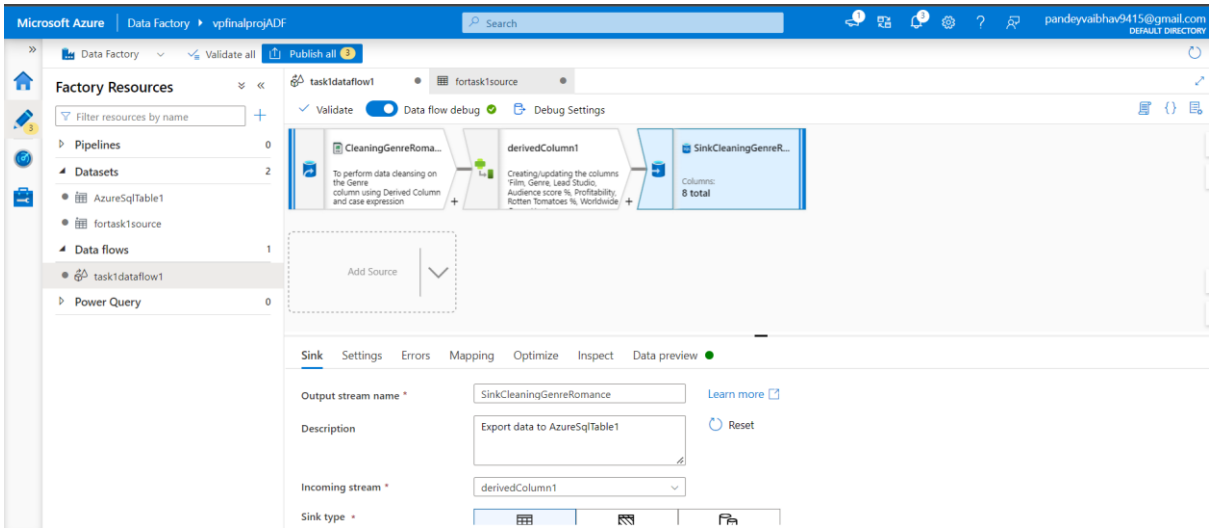
```
select * FROM movieoutput;
```

100 %

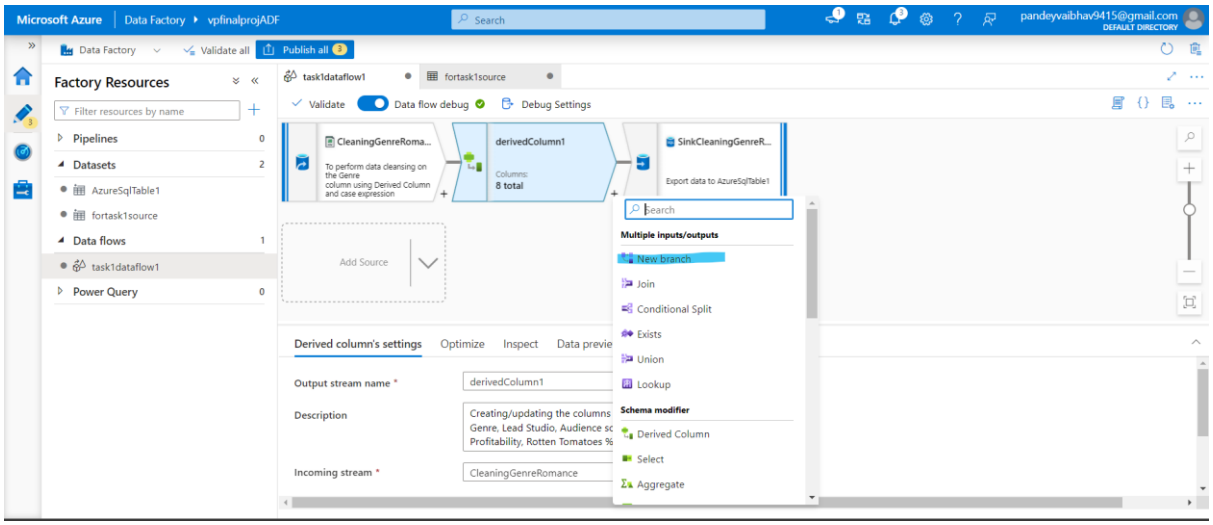
Results | Messages

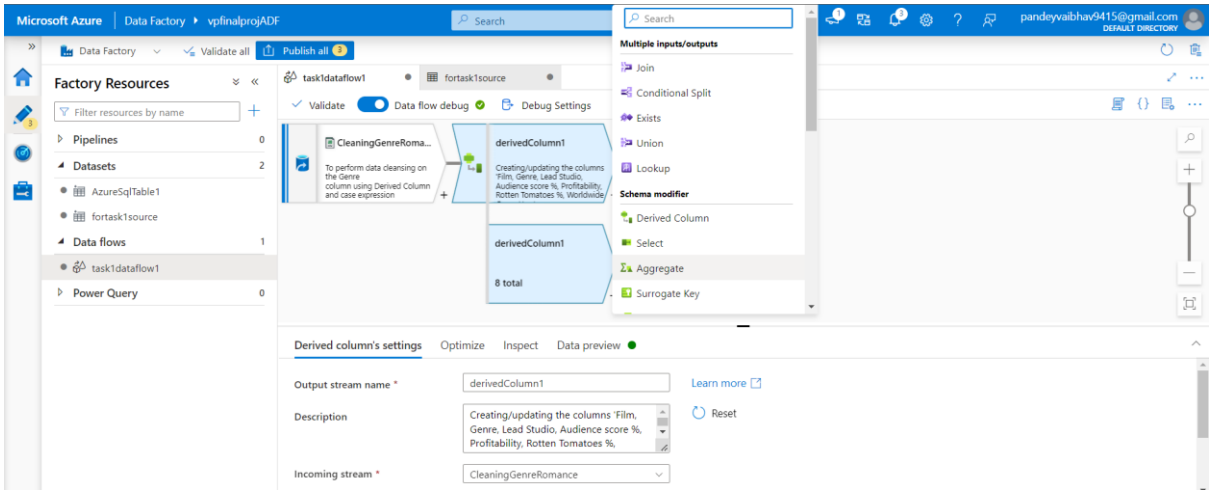
Film	Genre	Lead_Studio	Audience_score_prcnt	Profitability	Rotten_Tomatoes_prcnt	Worldwide_Gross	Year

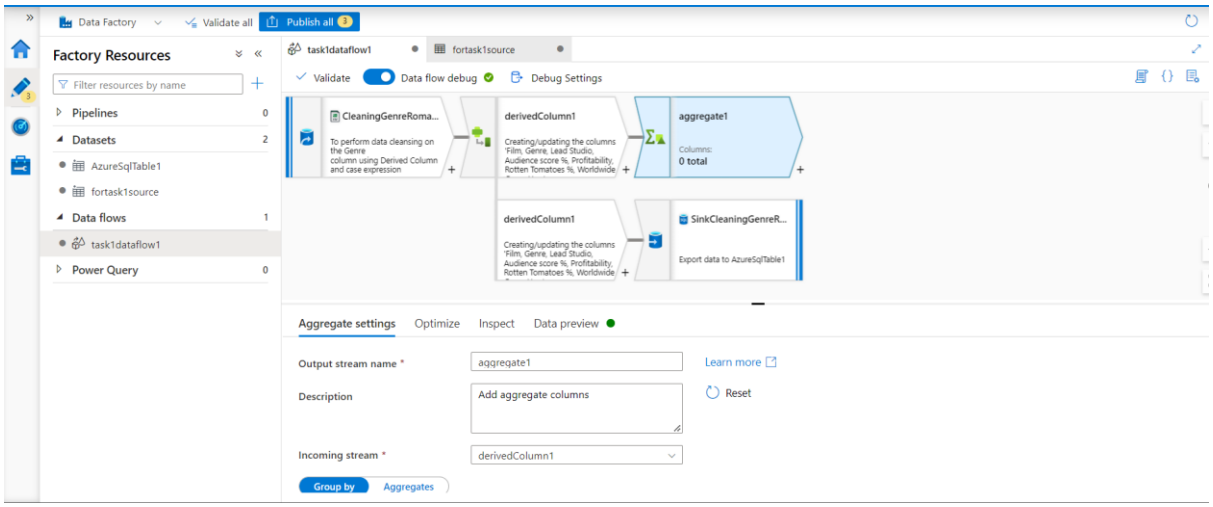


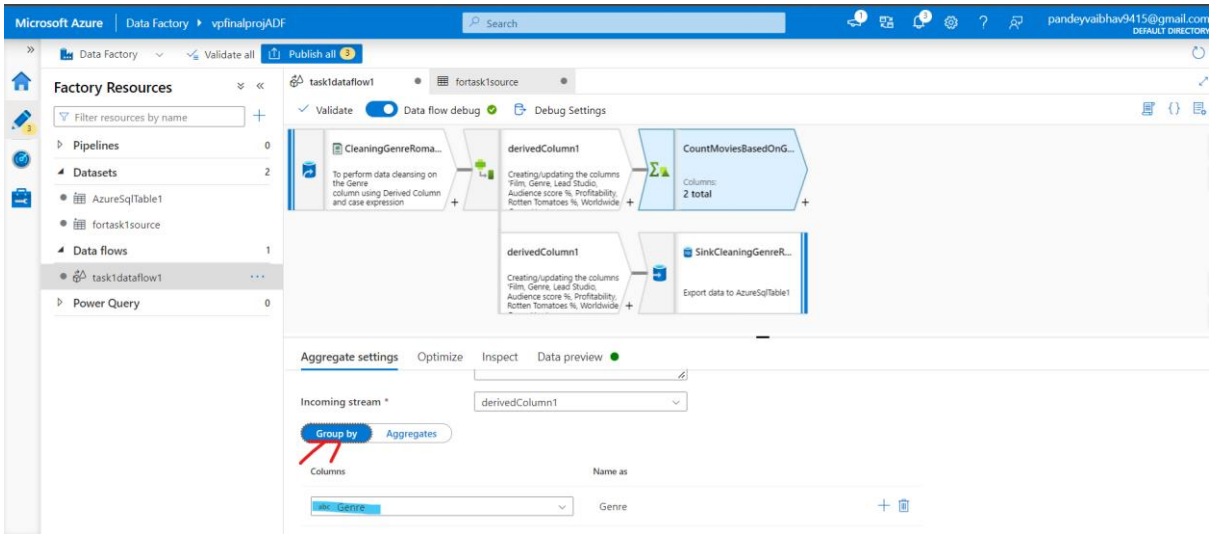


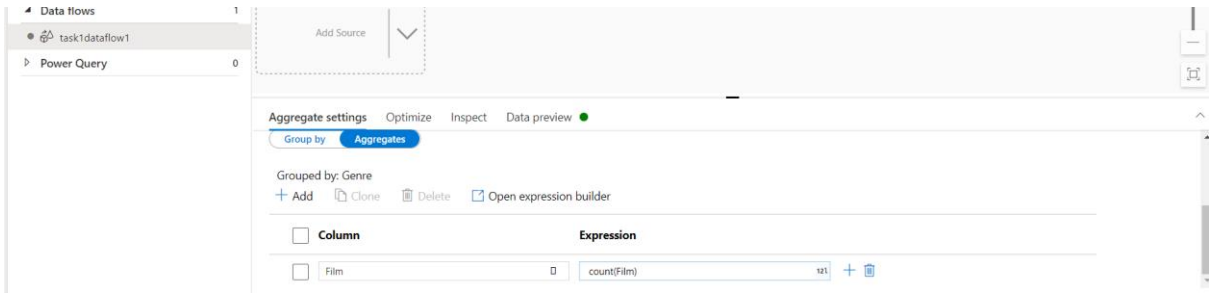
2. Create a data stream named CountMoviesBasedOnGenre that can calculate number of films for each genre and store it as a separate dataset in ADLS under folder name “solution/genreCount”

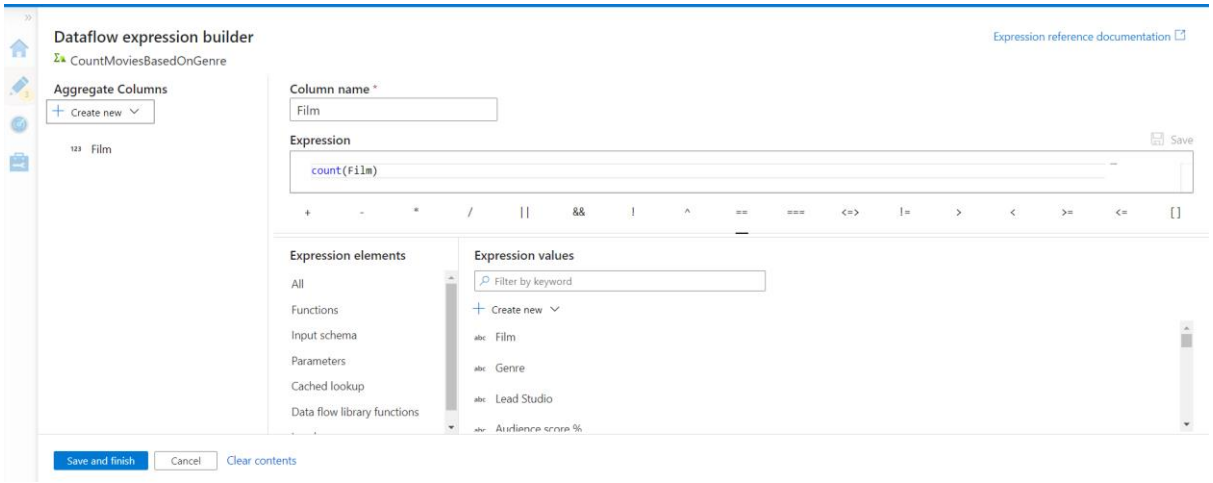


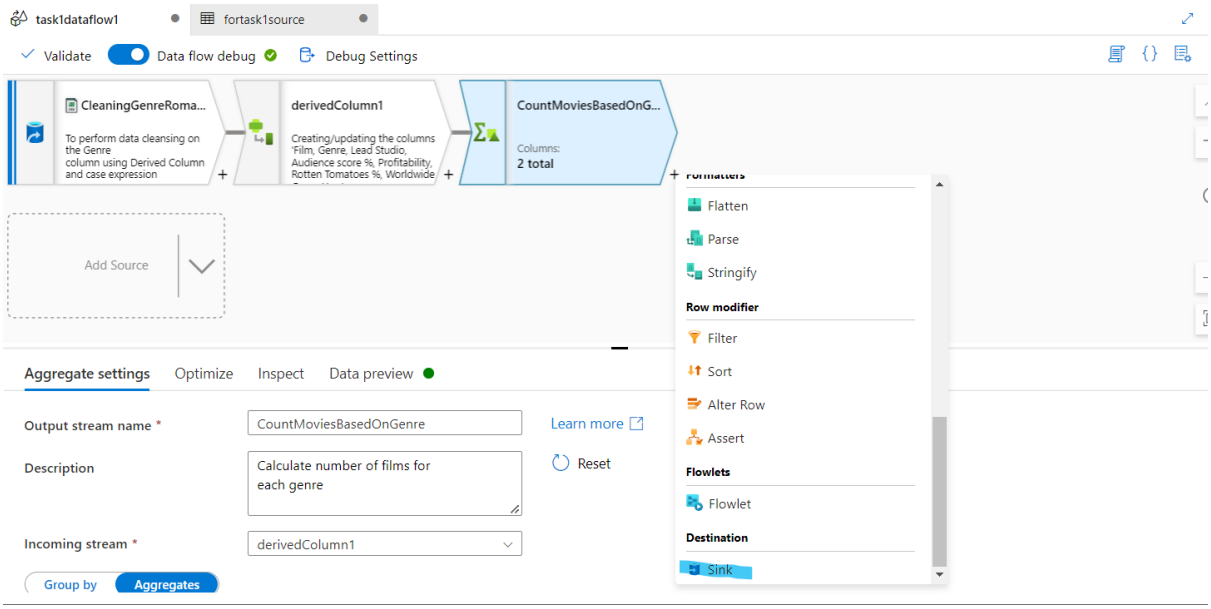


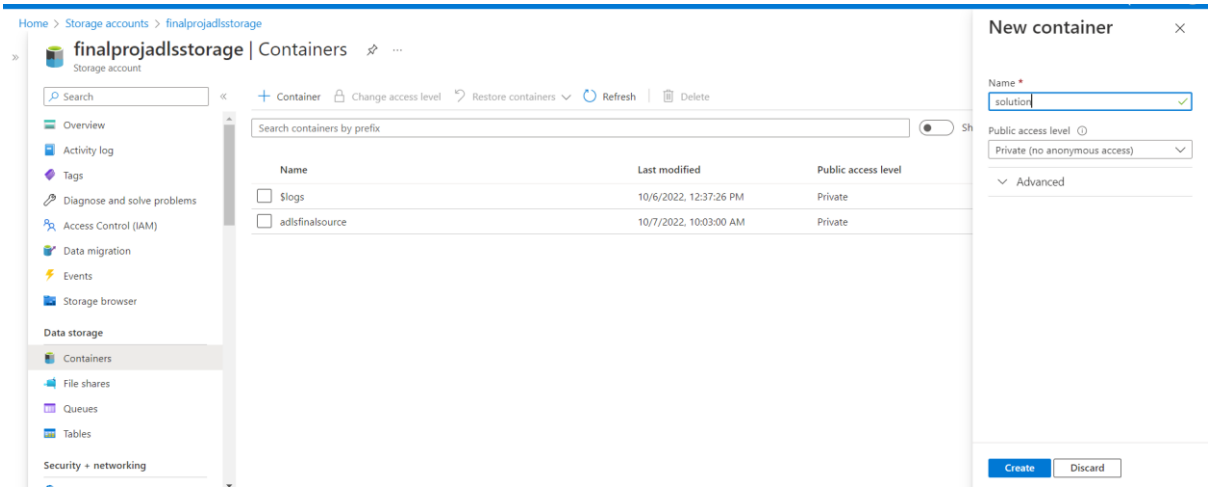


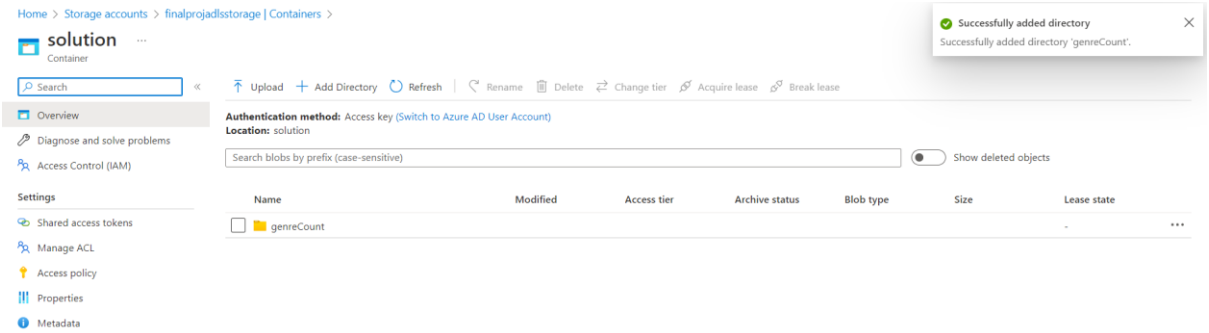

Microsoft Azure
Data Factory
vpfinalprojADF
Search
Validate all
Publish all
Factory Resources
Pipelines
Datasets
AzureSqlTable1
fortask1source
Data flows
task1dataflow1
Power Query
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
derivedColumn1
CountMoviesBasedOnG...
Columns:
2 total
derivedColumn1
SinkCleaningGenreR...
Export data to AzureSqlTable1
Aggregate settings
Optimize
Inspect
Data preview
Incoming stream *
derivedColumn1
Group by
Aggregates
Columns
Name as
Genre
Genre

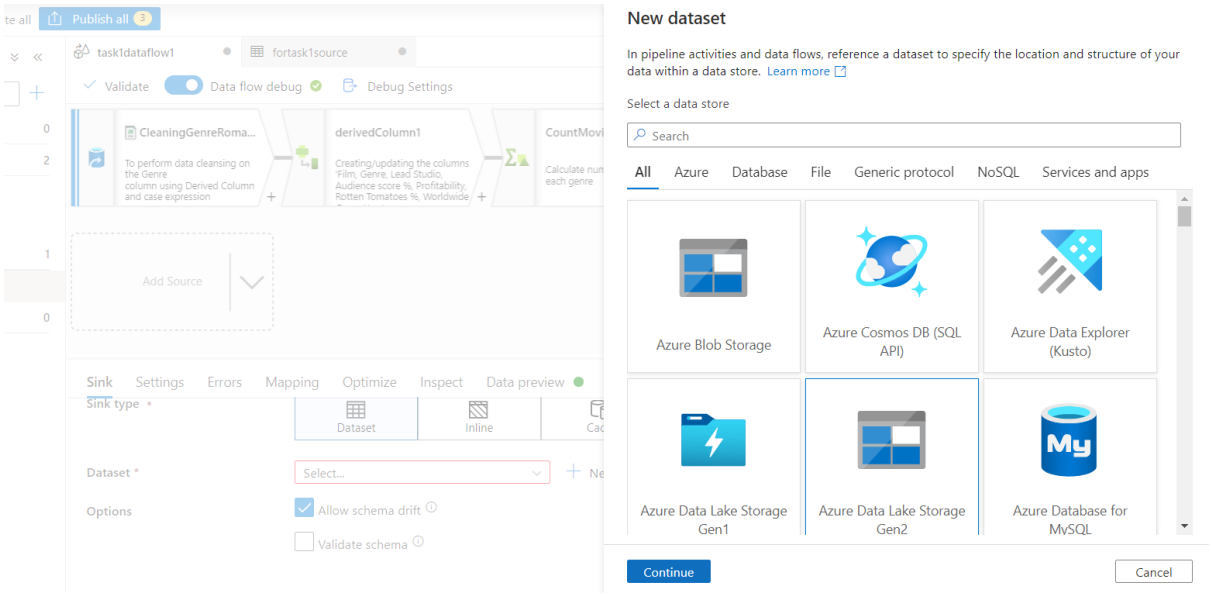

Data flows
task1dataflow1
Power Query
Add Source
Aggregate settings
Optimize
Inspect
Data preview
Group by
Aggregates
Grouped by: Genre
Add
Clone
Delete
Open expression builder
Column
Expression
Film
count(Film)

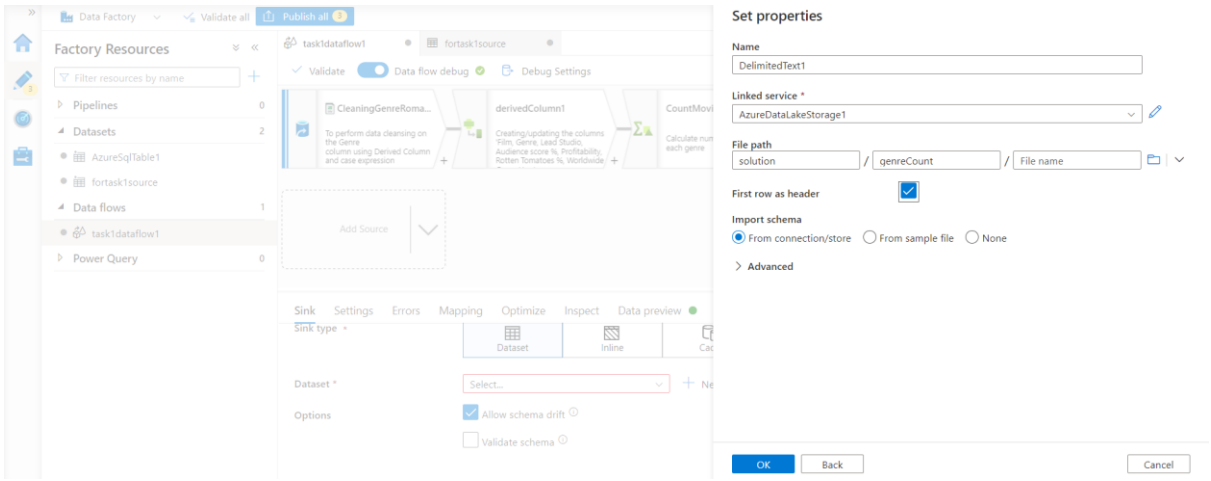

Dataflow expression builder
CountMoviesBasedOnGenre
Expression reference documentation
Aggregate Columns
Create new
Film
Column name *
Film
Expression
Save
count(Film)
+ - * / || && ! ^ == === <=> != > < >= <= []
Expression elements
All
Functions
Input schema
Parameters
Cached lookup
Data flow library functions
Expression values
Filter by keyword
Create new
Film
Genre
Lead Studio
Audience score %
Save and finish
Cancel
Clear contents

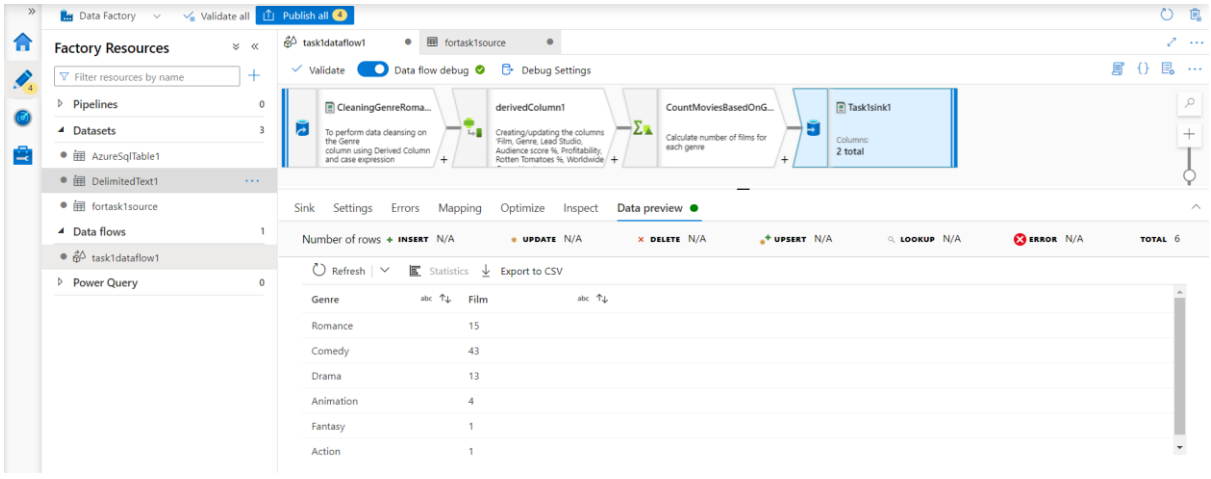

task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
To perform data cleansing on the Genre column using Derived Column and case expression
derivedColumn1
Creating/updating the columns 'Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide
CountMoviesBasedOnG...
Columns:
2 total
Add Source
Flatten
Parse
Stringify
Row modifier
Filter
Sort
Alter Row
Assert
Flowlets
Flowlet
Destination
Sink
Aggregate settings
Optimize
Inspect
Data preview
Output stream name *
CountMoviesBasedOnGenre
Learn more
Description
Calculate number of films for each genre
Reset
Incoming stream *
derivedColumn1
Group by
Aggregates

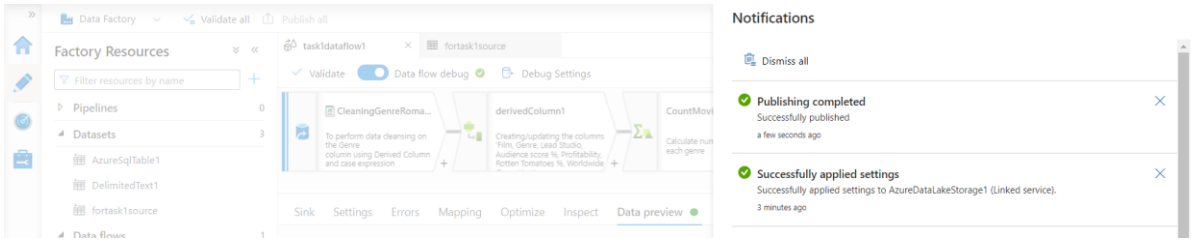

Home > Storage accounts > finalprojadlsstorage
finalprojadlsstorage | Containers
Storage account
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Data storage
Containers
File shares
Queues
Tables
Security + networking
Container
Change access level
Restore containers
Refresh
Delete
Search containers by prefix
Name
Last modified
Public access level
$logs
10/6/2022, 12:37:26 PM
Private
adlsfinalsource
10/7/2022, 10:03:00 AM
Private
New container
Name *
solution
Public access level
Private (no anonymous access)
Advanced
Create
Discard

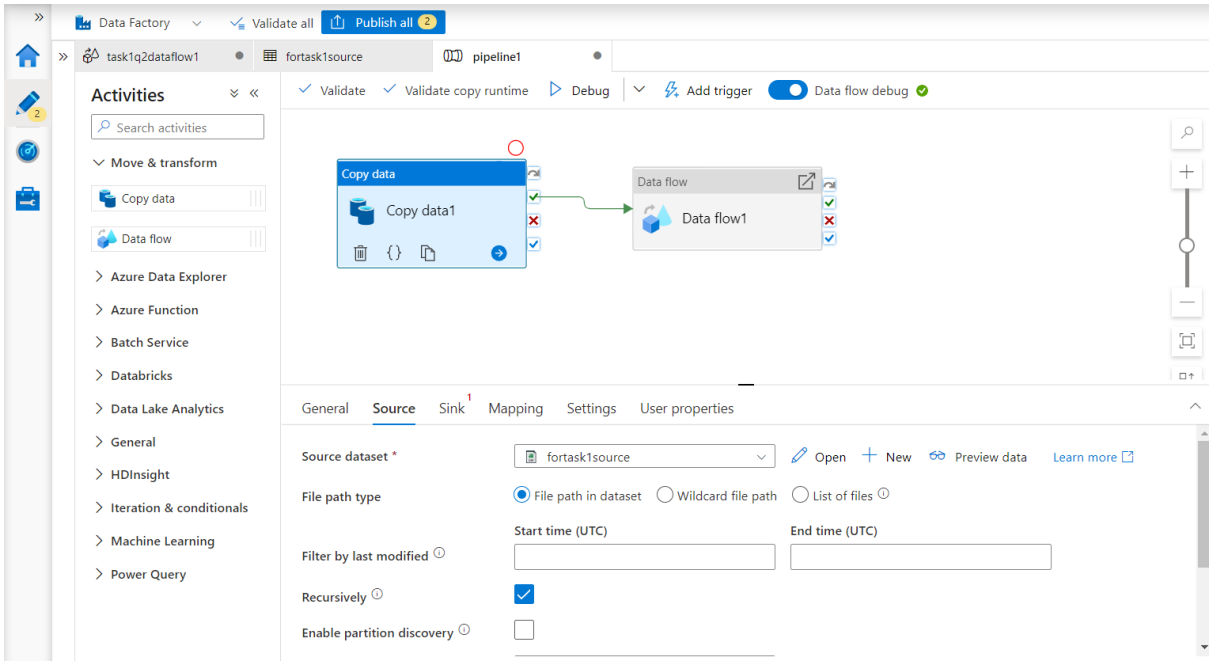

Home > Storage accounts > finalprojadlsstorage | Containers >
solution
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Authentication method: Access key (Switch to Azure AD User Account)
Location: solution
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
genreCount
Successfully added directory
Successfully added directory 'genreCount'.

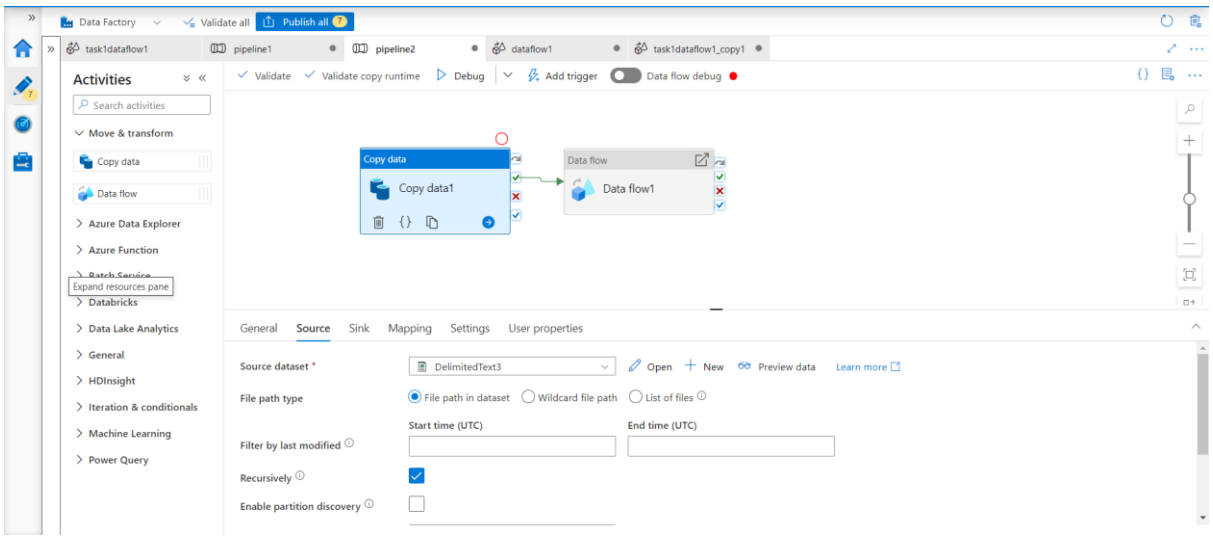

Publish all
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
derivedColumn1
Add Source
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Sink type
Dataset
Inline
Dataset *
Select...
Options
Allow schema drift
Validate schema
New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Blob Storage
Azure Cosmos DB (SQL API)
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2
Azure Database for MySQL
Continue
Cancel

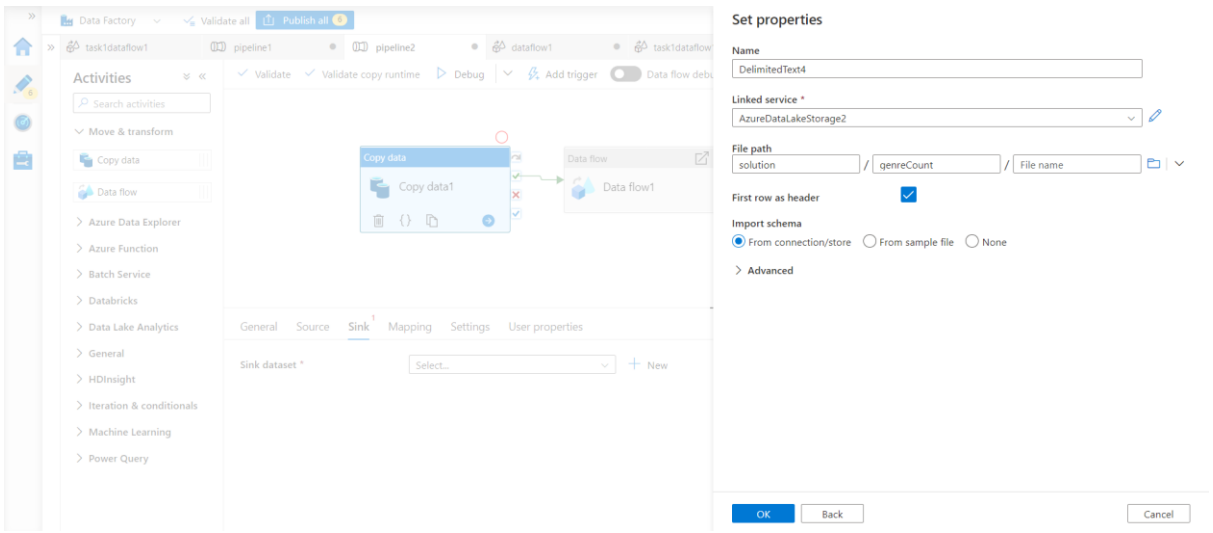

Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
Datasets
AzureSqlTable1
fortask1source
Data flows
task1dataflow1
Power Query
Set properties
Name
DelimitedText1
Linked service *
AzureDataLakeStorage1
File path
solution
genreCount
File name
First row as header
Import schema
From connection/store
From sample file
None
Advanced
OK
Back
Cancel

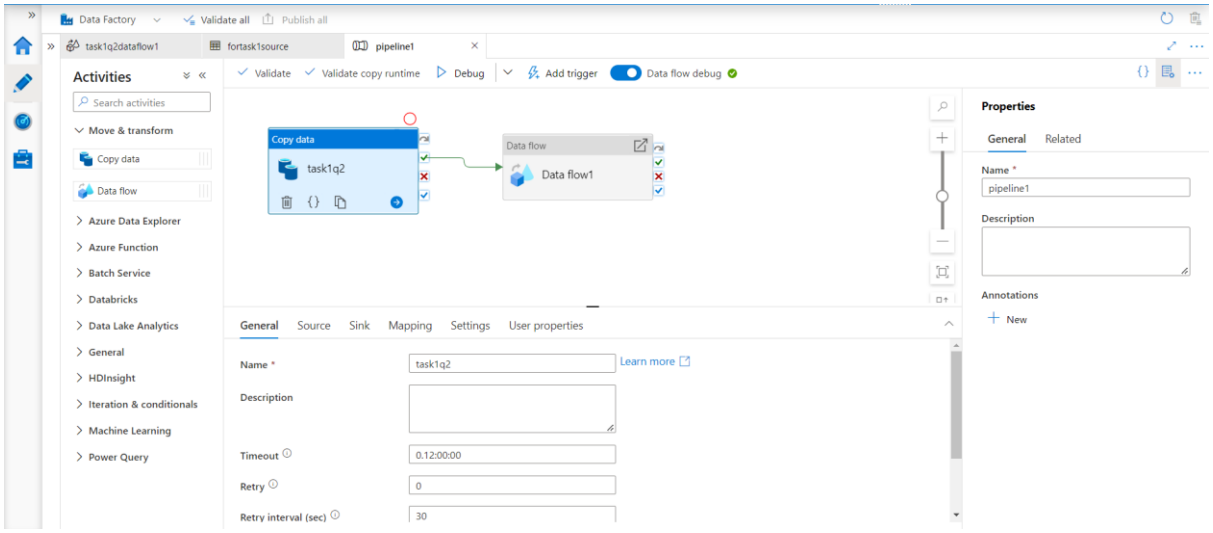

Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
Datasets
AzureSqlTable1
DelimitedText1
fortask1source
Data flows
task1dataflow1
Power Query
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
derivedColumn1
CountMoviesBasedOnG...
Calculate number of films for each genre
Task1sink1
Columns:
2 total
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Number of rows
INSERT N/A
UPDATE N/A
DELETE N/A
UPSERT N/A
LOOKUP N/A
ERROR N/A
TOTAL 6
Refresh
Statistics
Export to CSV
Genre
Film
Romance 15
Comedy 43
Drama 13
Animation 4
Fantasy 1
Action 1

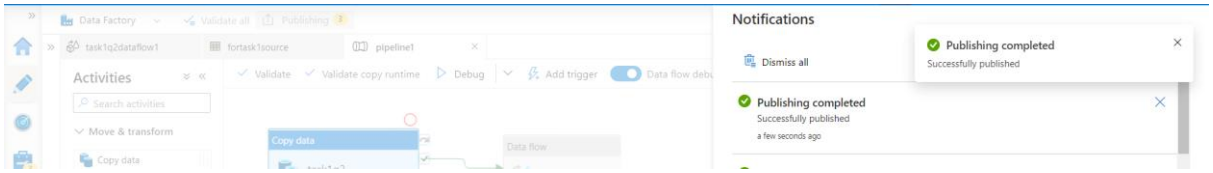

Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
Datasets
AzureSqlTable1
DelimitedText1
fortask1source
Data flows
task1dataflow1
Validate
Data flow debug
Debug Settings
CleaningGenreRoma...
derivedColumn1
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Notifications
Dismiss all
Publishing completed
Successfully published
a few seconds ago
Successfully applied settings
Successfully applied settings to AzureDataLakeStorage1 (Linked service).
3 minutes ago

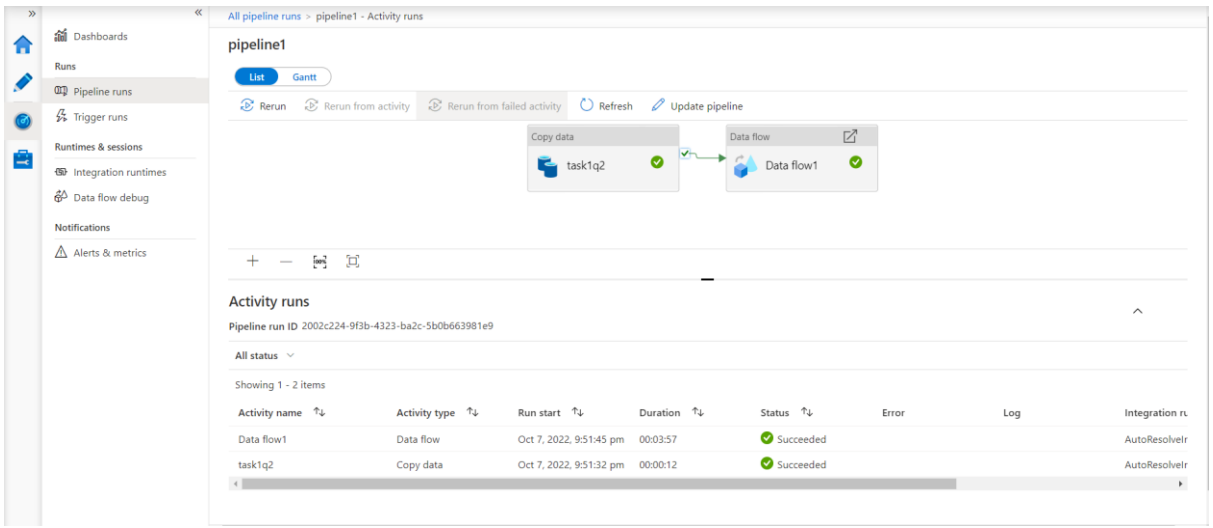

Data Factory
Validate all
Publish all 2
task1q2dataflow1
fortask1source
pipeline1
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Data flow debug
Copy data
Copy data1
Data flow
Data flow1
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
fortask1source
Open
New
Preview data
Learn more
File path type
File path in dataset
Wildcard file path
List of files
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively
Enable partition discovery

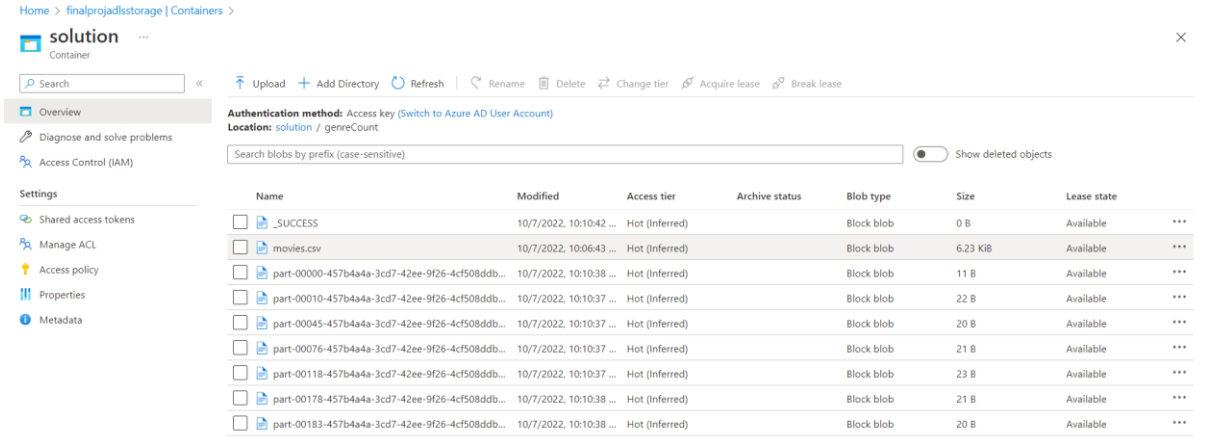

Data Factory
Validate all
Publish all 7
task1dataflow1
pipeline1
pipeline2
dataflow1
task1dataflow1_copy1
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Expand resources pane
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Data flow debug
Copy data
Copy data1
Data flow
Data flow1
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
DelimitedText3
Open
New
Preview data
Learn more
File path type
File path in dataset
Wildcard file path
List of files
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively
Enable partition discovery


Set properties
Name
DelimitedText4
Linked service *
AzureDataLakeStorage2
File path
solution
genreCount
File name
First row as header
Import schema
From connection/store
From sample file
None
Advanced
OK
Back
Cancel

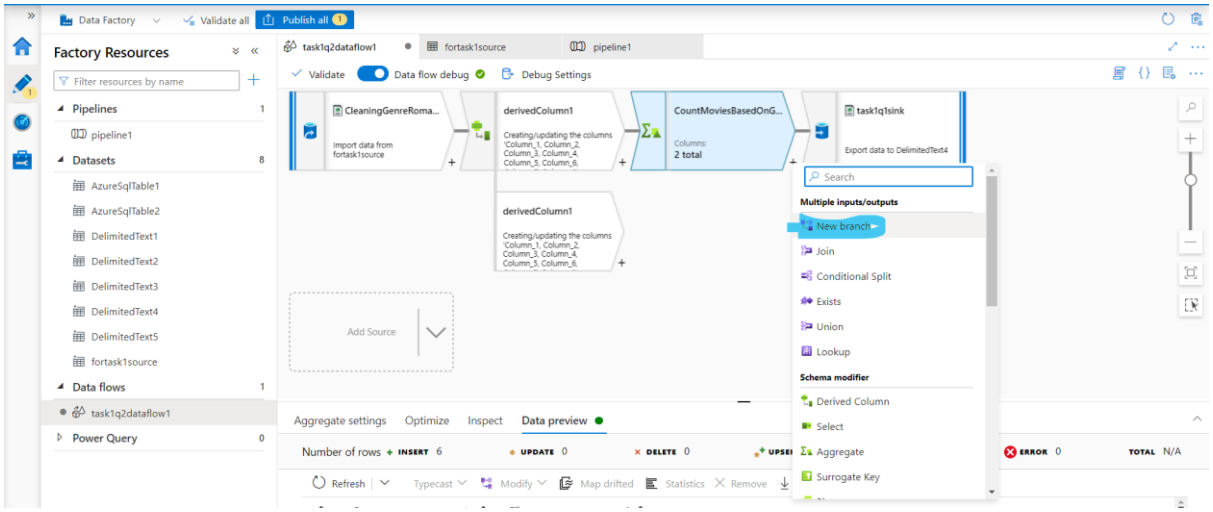

Data Factory
Validate all
Publish all
task1q2dataflow1
fortask1source
pipeline1
Activities
Search activities
Move & transform
Copy data
Data flow
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Data flow debug
Copy data
task1q2
Data flow
Data flow1
Properties
General
Related
Name *
pipeline1
Description
Annotations
New
General
Source
Sink
Mapping
Settings
User properties
Name *
task1q2
Learn more
Description
Timeout
0.12:00:00
Retry
0
Retry interval (sec)
30

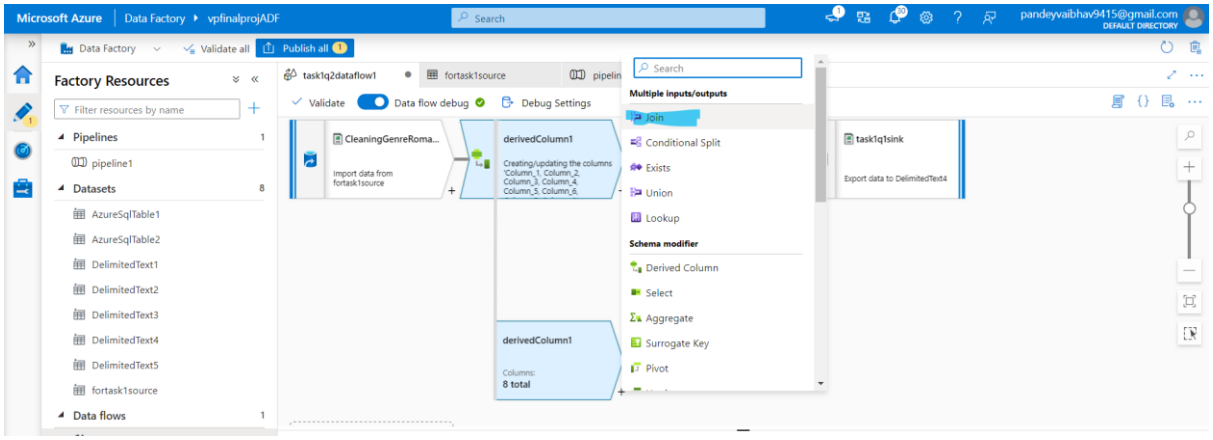

Notifications
Dismiss all
Publishing completed
Successfully published
Publishing completed
Successfully published
a few seconds ago

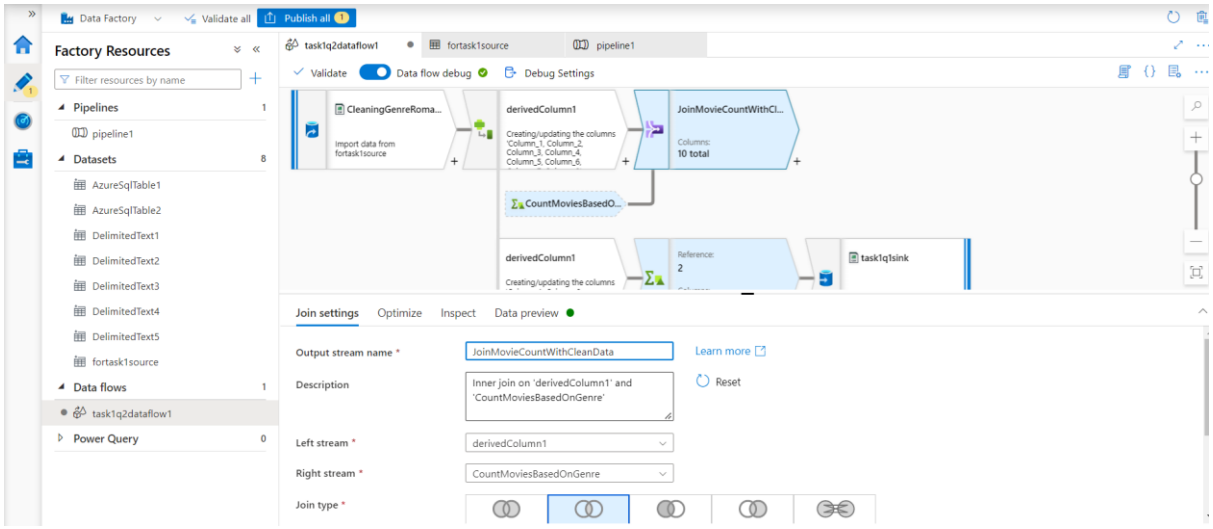

All pipeline runs > pipeline1 - Activity runs
Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
pipeline1
List
Gantt
Rerun
Rerun from activity
Rerun from failed activity
Refresh
Update pipeline
Copy data
task1q2
Data flow
Data flow1
Activity runs
Pipeline run ID 2002c224-9f3b-4323-ba2c-5b0b663981e9
All status
Showing 1 - 2 items
Activity name
Activity type
Run start
Duration
Status
Error
Log
Integration ru
Data flow1
Data flow
Oct 7, 2022, 9:51:45 pm
00:03:57
Succeeded
AutoResolveIr
task1q2
Copy data
Oct 7, 2022, 9:51:32 pm
00:00:12
Succeeded
AutoResolveIr

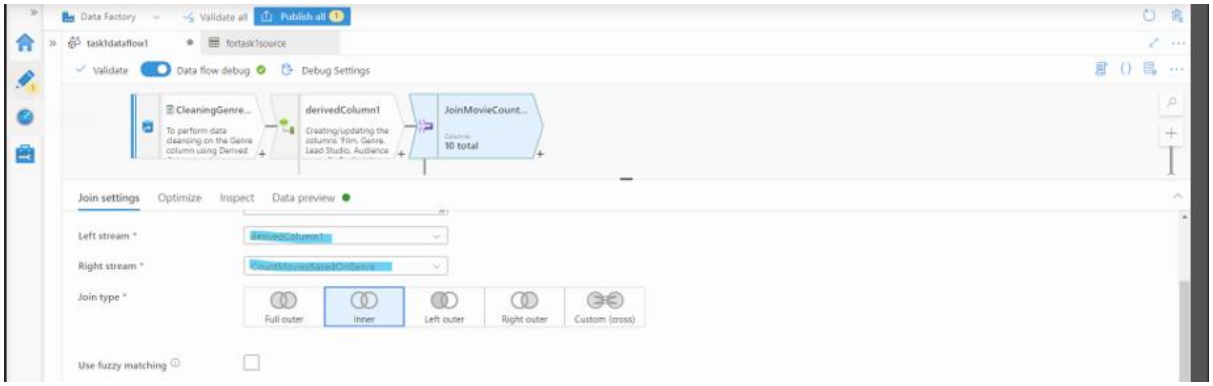

Home > finalprojadlsstorage | Containers >
solution
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Authentication method: Access key (Switch to Azure AD User Account)
Location: solution / genreCount
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
_SUCCESS
movies.csv
part-00000-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00010-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00045-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00076-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00118-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00178-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
part-00183-457b4a4a-3cd7-42ee-9f26-4cf508ddb...
10/7/2022, 10:10:42 ...
10/7/2022, 10:06:43 ...
10/7/2022, 10:10:38 ...
10/7/2022, 10:10:37 ...
Hot (Inferred)
Block blob
0 B
6.23 KiB
11 B
22 B
20 B
21 B
23 B
Available

3. Create a new stream named JoinMovieCountWithCleanData. Perform join operation on CountMoviesBasedOnGenre with CleaningGenreRomance stream and store the same in the Azure SQL Database.

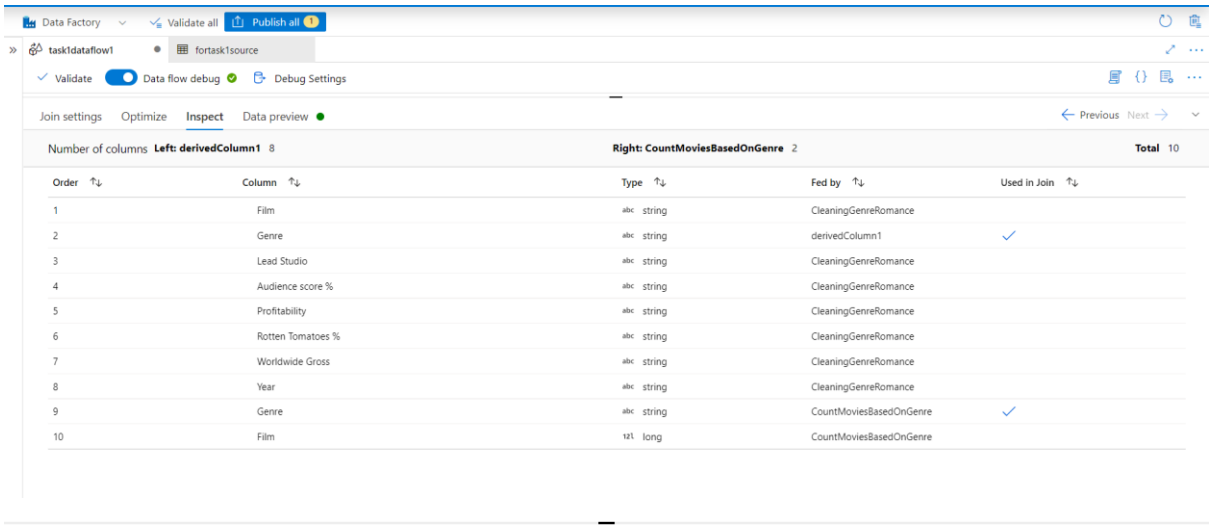


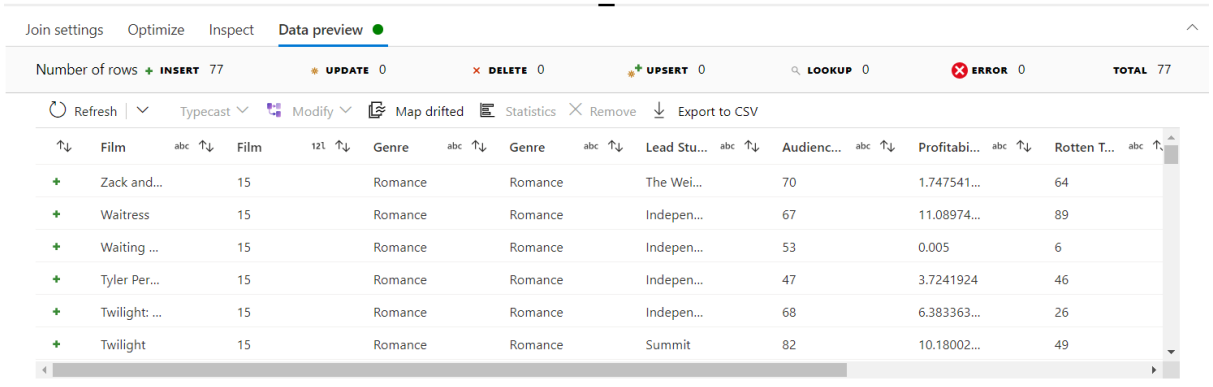


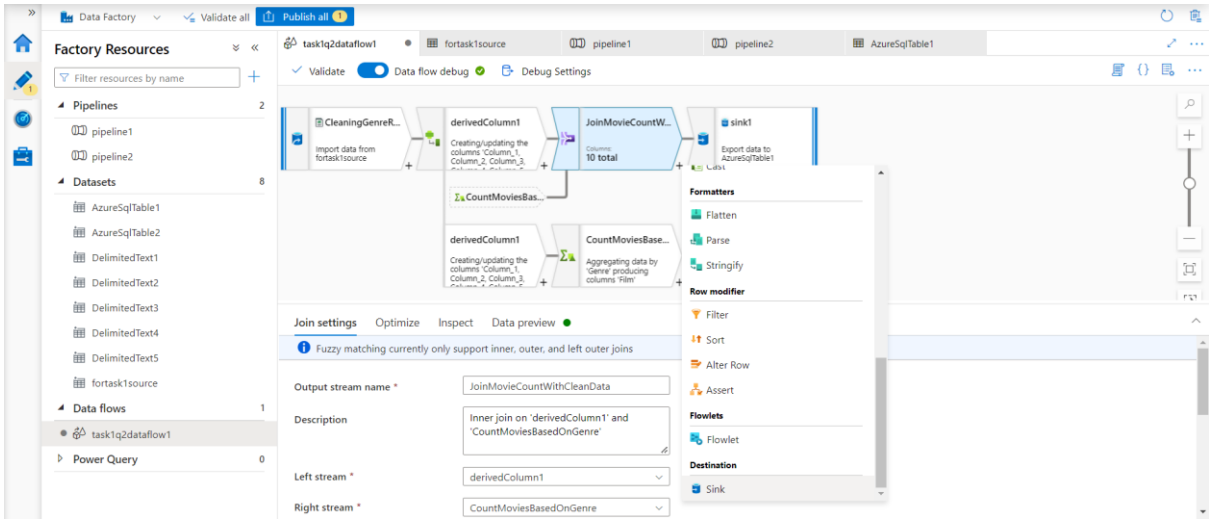


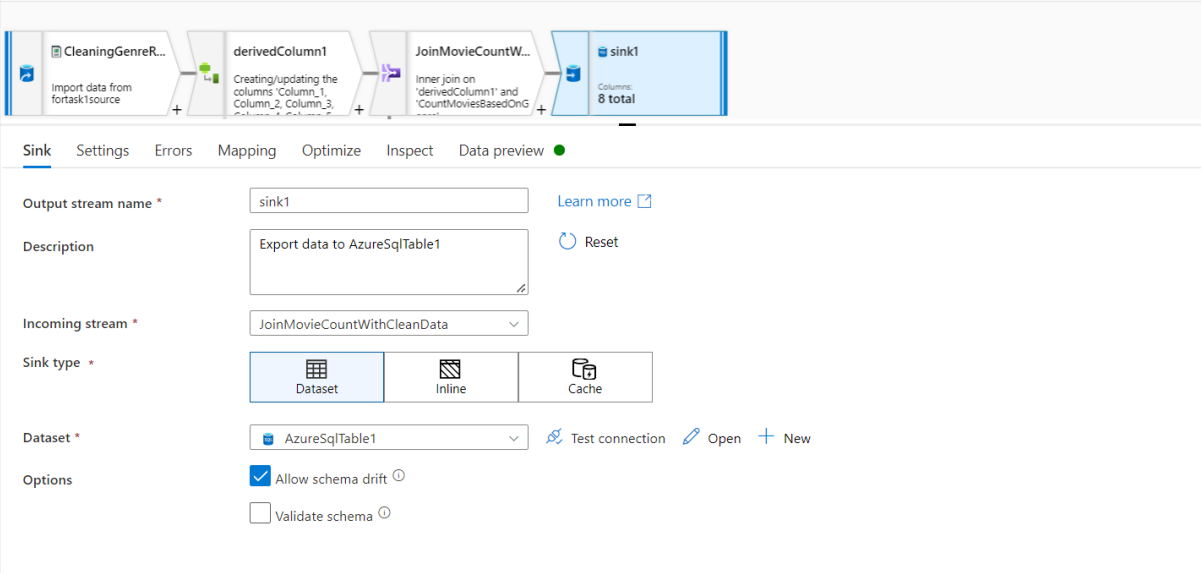

Data Factory
Validate all
Publish all
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenre...
derivedColumn1
JoinMovieCount...
10 total
Join settings
Optimize
Inspect
Data preview
Left stream *
Right stream *
Join type *
Full outer
Inner
Left outer
Right outer
Custom (cross)
Use fuzzy matching

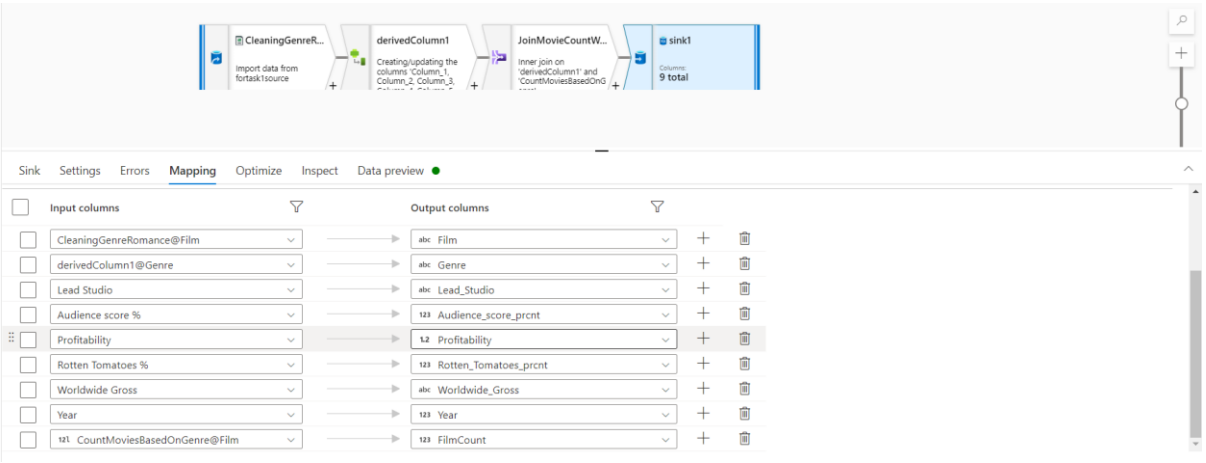

Join settings
Optimize
Inspect
Data preview
Join type *
Full outer
Inner
Left outer
Right outer
Custom (cross)
Use fuzzy matching
Combine text parts
Similarity score column
Join conditions *
Left: derivedColumn1's column
Similarity Threshold
Right: CountMoviesBasedOnGenre's column
Genre
80
Genre

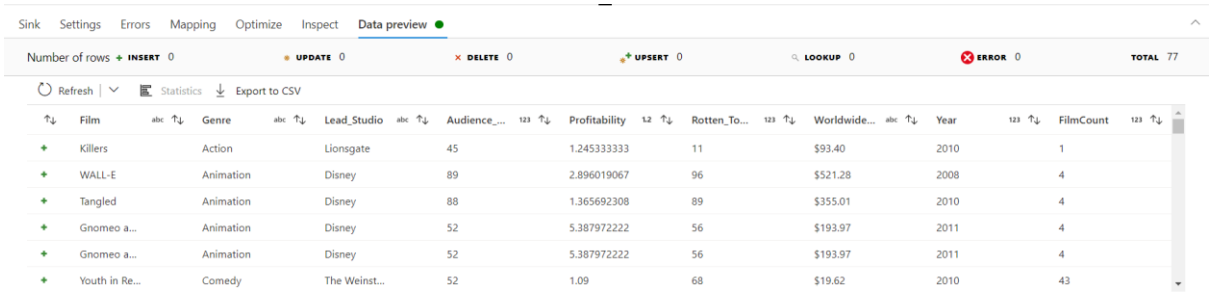

Data Factory
Validate all
Publish all
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
Join settings
Optimize
Inspect
Data preview
Previous
Next
Number of columns Left: derivedColumn1 8
Right: CountMoviesBasedOnGenre 2
Total 10
Order
Column
Type
Fed by
Used in Join
1 Film string CleaningGenreRomance
2 Genre string derivedColumn1
3 Lead Studio string CleaningGenreRomance
4 Audience score % string CleaningGenreRomance
5 Profitability string CleaningGenreRomance
6 Rotten Tomatoes % string CleaningGenreRomance
7 Worldwide Gross string CleaningGenreRomance
8 Year string CleaningGenreRomance
9 Genre string CountMoviesBasedOnGenre
10 Film long CountMoviesBasedOnGenre

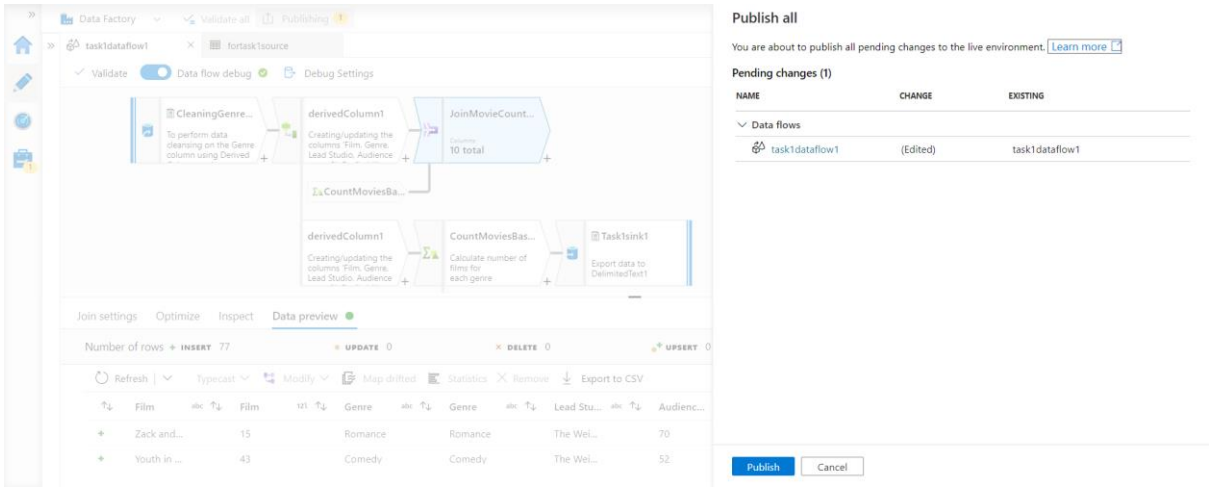

Join settings
Optimize
Inspect
Data preview
Number of rows INSERT 77
UPDATE 0
DELETE 0
UPSERT 0
LOOKUP 0
ERROR 0
TOTAL 77
Refresh
Typecast
Modify
Map drifted
Statistics
Remove
Export to CSV
Film
Film
Genre
Genre
Lead Stu...
Audienc...
Profitabi...
Rotten T...
Zack and... 15 Romance Romance The Wei... 70 1.747541... 64
Waitress 15 Romance Romance Indepen... 67 11.08974... 89
Waiting ... 15 Romance Romance Indepen... 53 0.005 6
Tyler Per... 15 Romance Romance Indepen... 47 3.7241924 46
Twilight: ... 15 Romance Romance Indepen... 68 6.383363... 26
Twilight 15 Romance Romance Summit 82 10.18002... 49

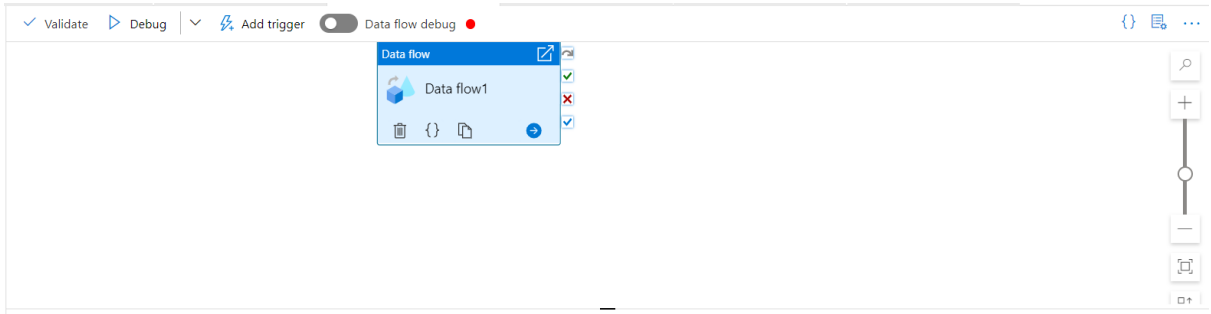

Data Factory
Validate all
Publish all
Factory Resources
Filter resources by name
Pipelines
pipeline1
pipeline2
Datasets
AzureSqlTable1
AzureSqlTable2
DelimitedText1
DelimitedText2
DelimitedText3
DelimitedText4
DelimitedText5
fortask1source
Data flows
task1q2dataflow1
Power Query
task1q2dataflow1
fortask1source
pipeline1
pipeline2
AzureSqlTable1
Validate
Data flow debug
Debug Settings
CleaningGenreR...
Import data from fortask1source
derivedColumn1
Creating/updating the columns 'Column_1, Column_2, Column_3,
JoinMovieCountW...
10 total
sink1
Export data to AzureSqlTable1
CountMoviesBas...
CountMoviesBase...
Aggregating data by 'Genre' producing columns 'Film'
Formatters
Flatten
Parse
Stringify
Row modifier
Filter
Sort
Alter Row
Assert
Flowlets
Flowlet
Destination
Sink
Join settings
Optimize
Inspect
Data preview
Fuzzy matching currently only support inner, outer, and left outer joins
Output stream name *
JoinMovieCountWithCleanData
Description
Inner join on 'derivedColumn1' and 'CountMoviesBasedOnGenre'
Left stream *
derivedColumn1
Right stream *
CountMoviesBasedOnGenre

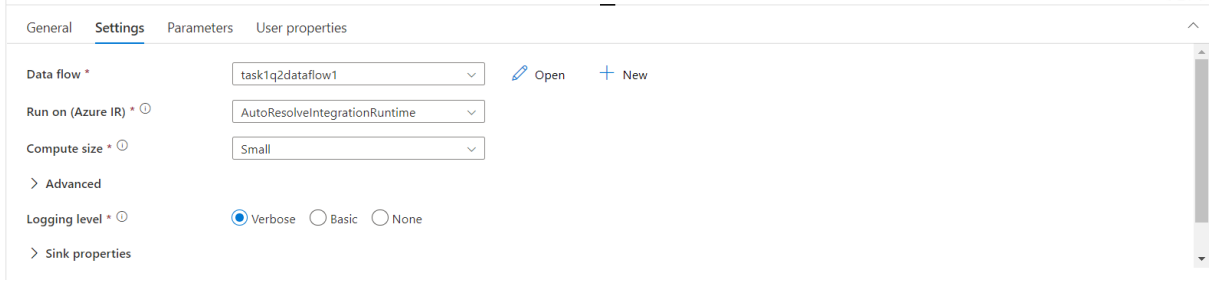

CleaningGenreR...
Import data from fortask1source
derivedColumn1
Creating/updating the columns 'Column_1, Column_2, Column_3,
JoinMovieCountW...
Inner join on 'derivedColumn1' and 'CountMoviesBasedOnG
sink1
Columns:
8 total
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Output stream name *
sink1
Learn more
Description
Export data to AzureSqlTable1
Reset
Incoming stream *
JoinMovieCountWithCleanData
Sink type *
Dataset
Inline
Cache
Dataset *
AzureSqlTable1
Test connection
Open
New
Options
Allow schema drift
Validate schema

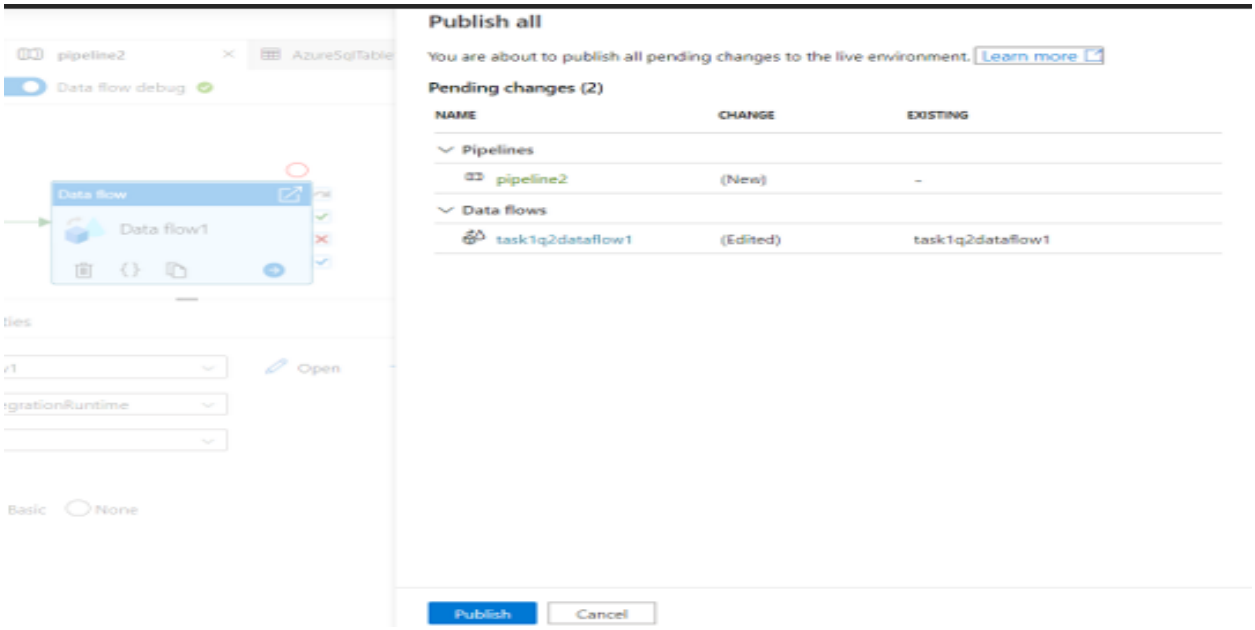

CleaningGenreR...
Import data from fortask1source
derivedColumn1
Creating/updating the columns 'Column_1, Column_2, Column_3,
JoinMovieCountW...
Inner join on 'derivedColumn1' and 'CountMoviesBasedOnG
sink1
Columns:
9 total
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Input columns
Output columns
CleaningGenreRomance@Film
Film
derivedColumn1@Genre
Genre
Lead Studio
Lead_Studio
Audience score %
Audience_score_prcnt
Profitability
Profitability
Rotten Tomatoes %
Rotten_Tomatoes_prcnt
Worldwide Gross
Worldwide_Gross
Year
Year
CountMoviesBasedOnGenre@Film
FilmCount

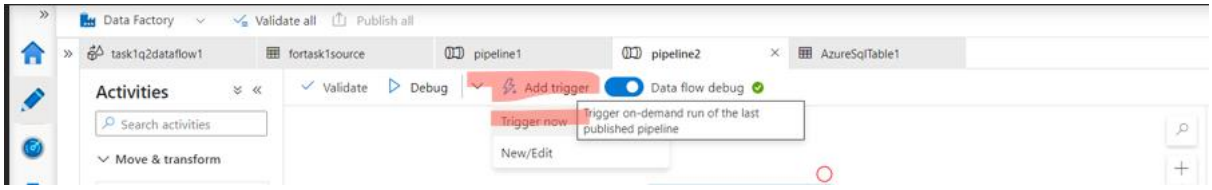

Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Number of rows + INSERT 0
UPDATE 0
DELETE 0
UPSERT 0
LOOKUP 0
ERROR 0
TOTAL 77
Refresh
Statistics
Export to CSV
Film
Genre
Lead_Studio
Audience_...
Profitability
Rotten_To...
Worldwide...
Year
FilmCount
Killers
Action
Lionsgate
45
1.245333333
11
$93.40
2010
1
WALL-E
Animation
Disney
89
2.896019067
96
$521.28
2008
4
Tangled
Animation
Disney
88
1.365692308
89
$355.01
2010
4
Gnomeo a...
Animation
Disney
52
5.387972222
56
$193.97
2011
4
Gnomeo a...
Animation
Disney
52
5.387972222
56
$193.97
2011
4
Youth in Re...
Comedy
The Weinst...
52
1.09
68
$19.62
2010
43

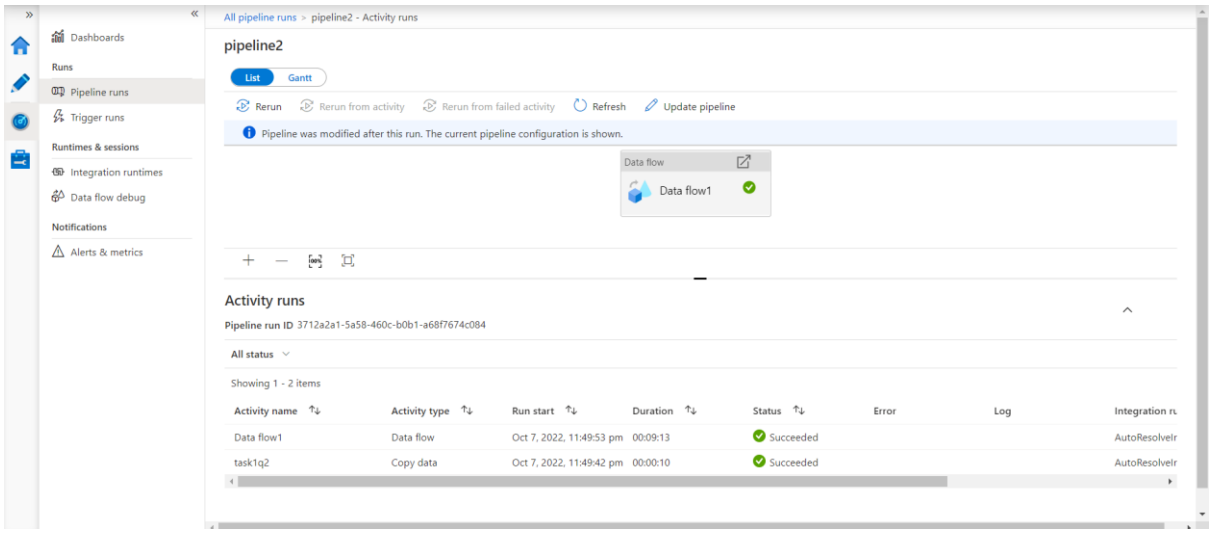

Data Factory
Validate all
Publishing
task1dataflow1
fortask1source
Validate
Data flow debug
Debug Settings
CleaningGenre...
derivedColumn1
JoinMovieCount...
10 total
CountMoviesBa...
derivedColumn1
CountMoviesBas...
Task1sink1
Join settings
Optimize
Inspect
Data preview
Number of rows + INSERT 77
UPDATE 0
DELETE 0
UPSERT 0
Refresh
Typecast
Modify
Map drifted
Statistics
Remove
Export to CSV
Film
Film
Genre
Genre
Lead Stu...
Audienc...
Zack and...
15
Romance
Romance
The Wei...
70
Youth in ...
43
Comedy
Comedy
The Wei...
52
Publish all
You are about to publish all pending changes to the live environment.
Learn more
Pending changes (1)
NAME
CHANGE
EXISTING
Data flows
task1dataflow1
(Edited)
task1dataflow1
Publish
Cancel

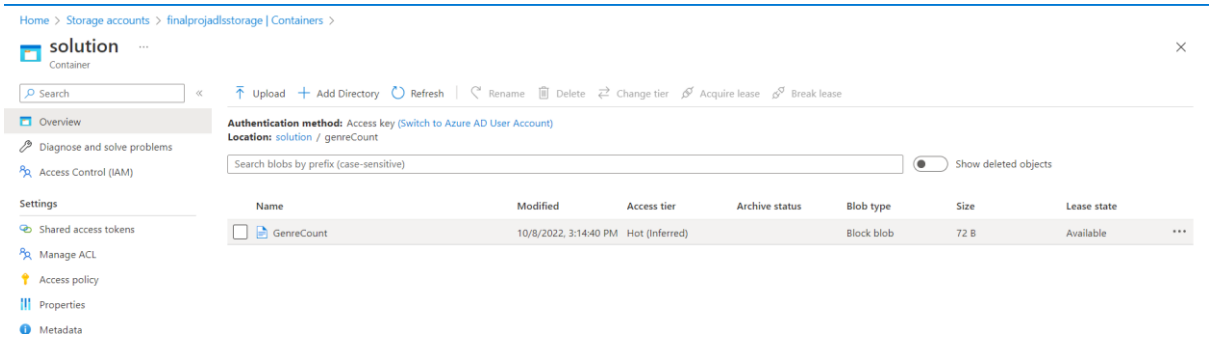

Validate
Debug
Add trigger
Data flow debug
Data flow
Data flow1

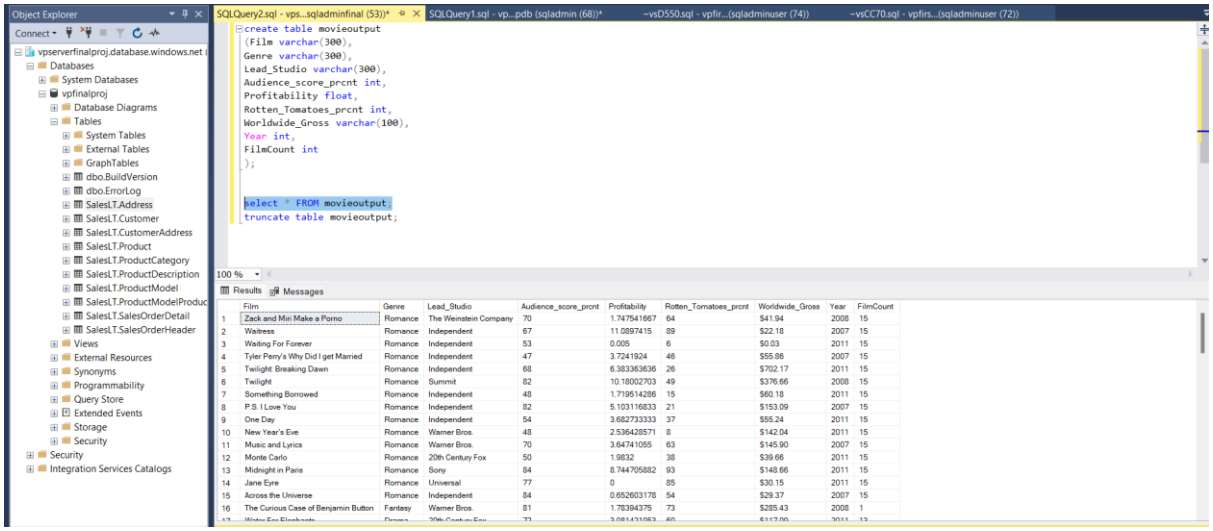

General
Settings
Parameters
User properties
Data flow *
task1q2dataflow1
Open
New
Run on (Azure IR) *
AutoResolveIntegrationRuntime
Compute size *
Small
Advanced
Logging level *
Verbose
Basic
None
Sink properties


Publish all
You are about to publish all pending changes to the live environment. Learn more
Pending changes (2)
NAME
CHANGE
EXISTING
Pipelines
pipeline2
(New)
Data flows
task1q2dataflow1
(Edited)
task1q2dataflow1
Publish
Cancel


Data Factory
Validate all
Publish all
task1q2dataflow1
fortask1source
pipeline1
pipeline2
AzureSqlTable1
Activities
Search activities
Move & transform
Validate
Debug
Add trigger
Data flow debug
Trigger now
Trigger on-demand run of the last published pipeline
New/Edit

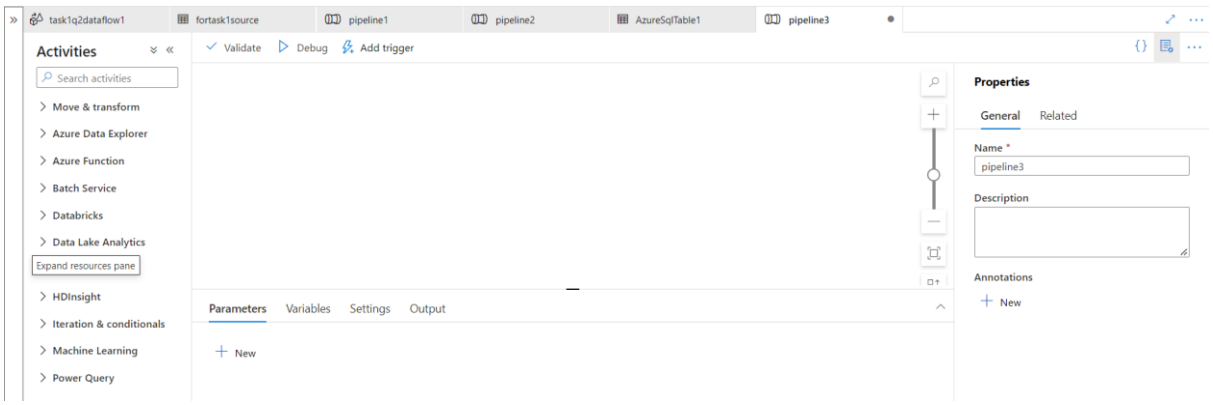

All pipeline runs > pipeline2 - Activity runs
pipeline2
List
Gantt
Rerun
Rerun from activity
Rerun from failed activity
Refresh
Update pipeline
Pipeline was modified after this run. The current pipeline configuration is shown.
Data flow
Data flow1
Activity runs
Pipeline run ID 3712a2a1-5a58-460c-b0b1-a68f7674c084
All status
Showing 1 - 2 items
Activity name
Activity type
Run start
Duration
Status
Error
Log
Data flow1
Data flow
Oct 7, 2022, 11:49:53 pm
00:09:13
Succeeded
task1q2
Copy data
Oct 7, 2022, 11:49:42 pm
00:00:10
Succeeded
Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics

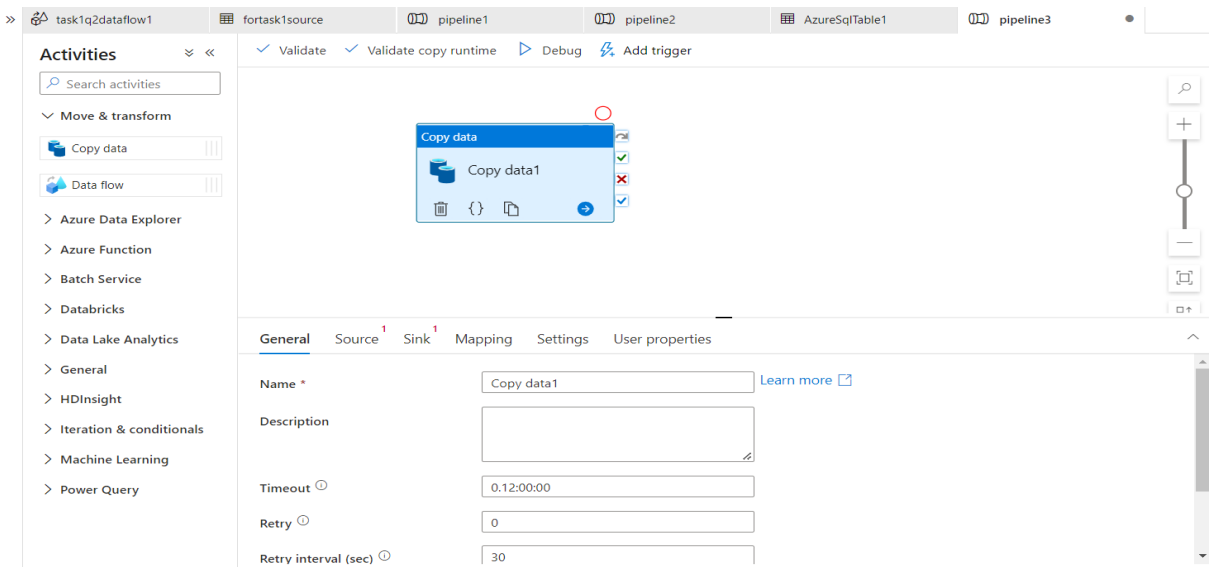

Home > Storage accounts > finalprojadlsstorage | Containers >
solution
Container
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Authentication method: Access key (Switch to Azure AD User Account)
Location: solution / genreCount
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
GenreCount
10/8/2022, 3:14:40 PM
Hot (Inferred)
Block blob
72 B
Available

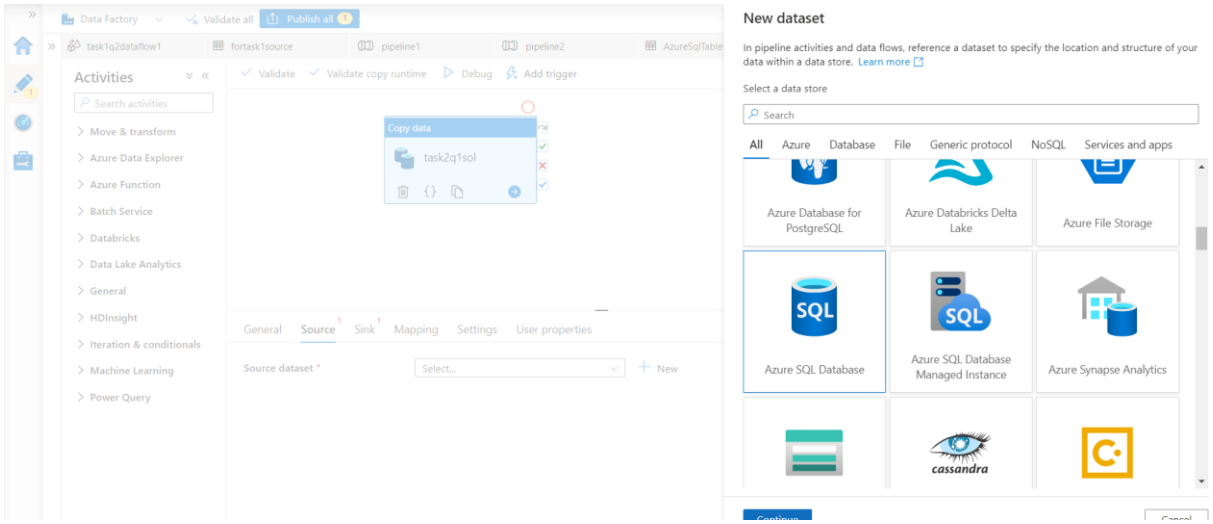


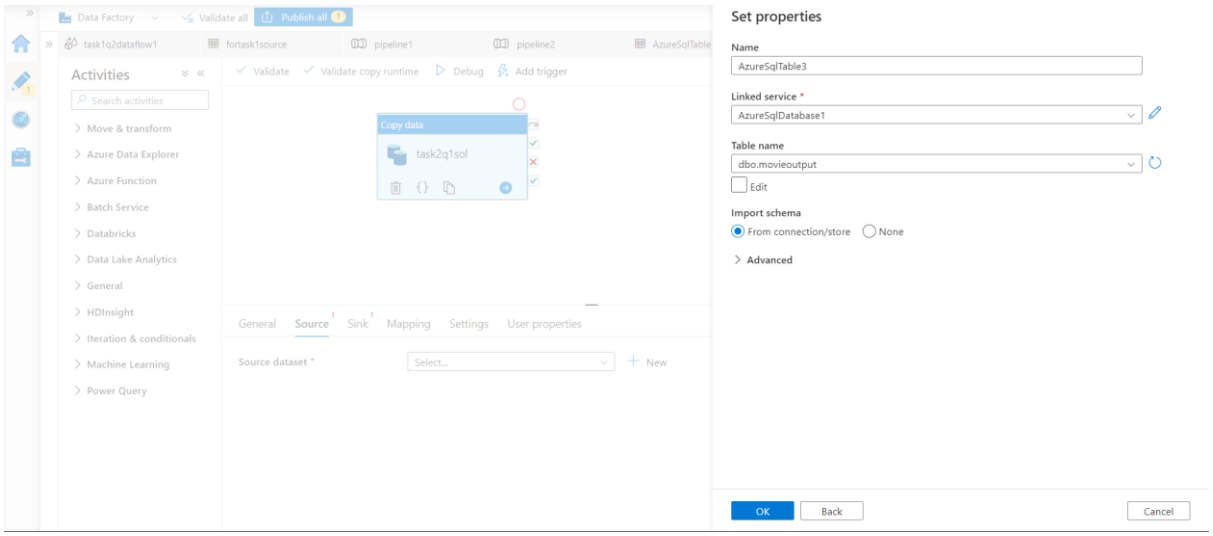
Task 2: Create the following activity pipeline

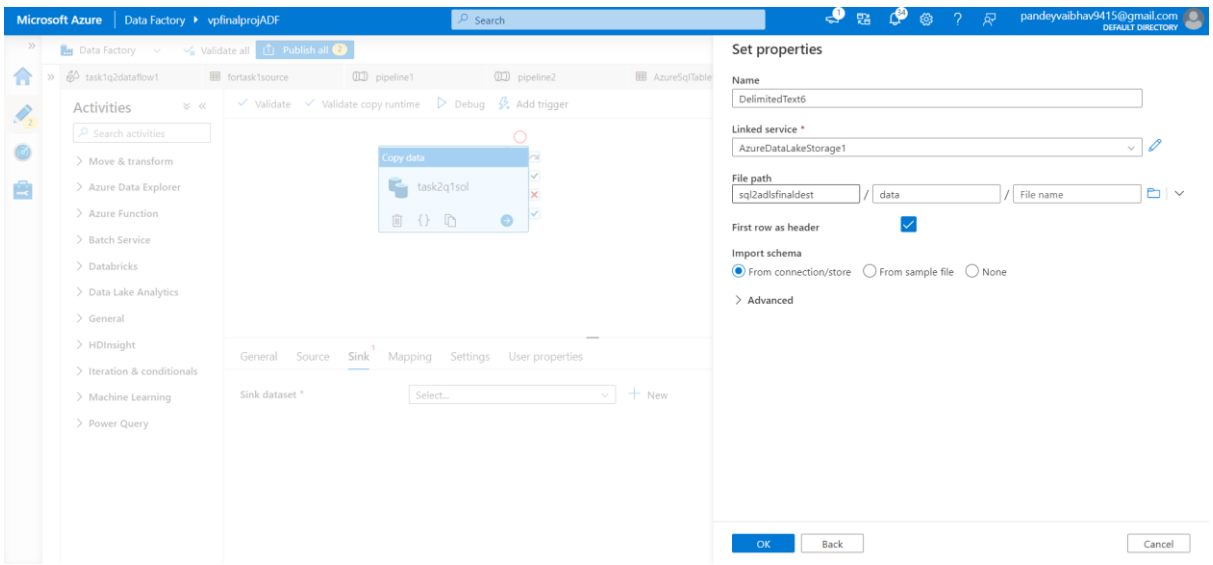
1. Get the clean data from Azure SQL DB. Create an activity that can copy the data from SQLDB to ADLS Gen2.

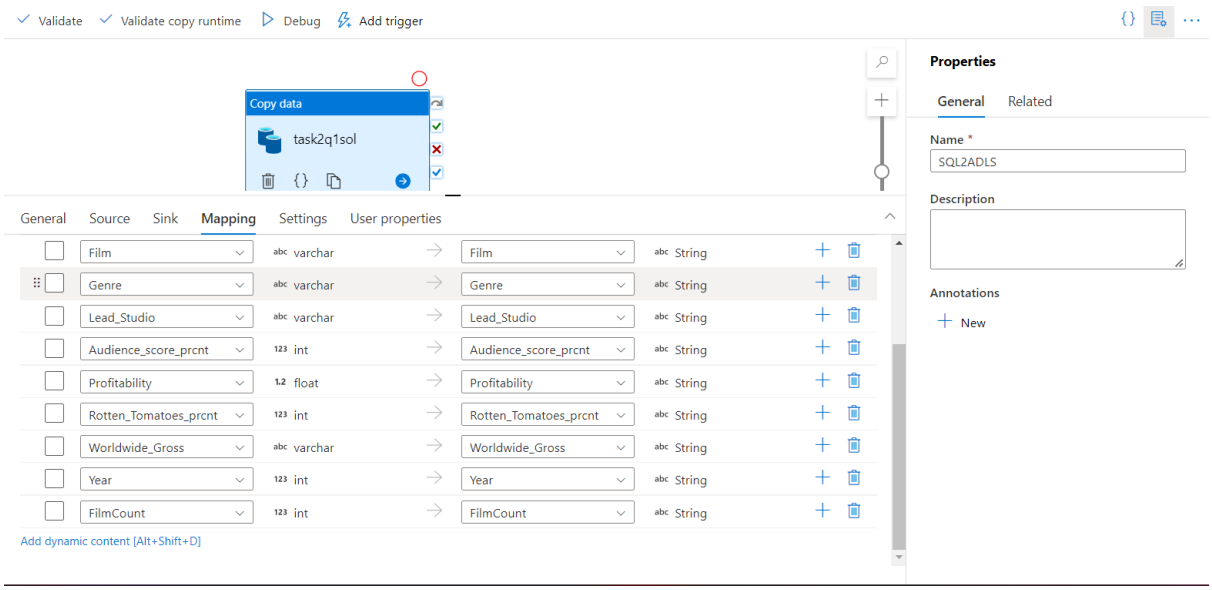


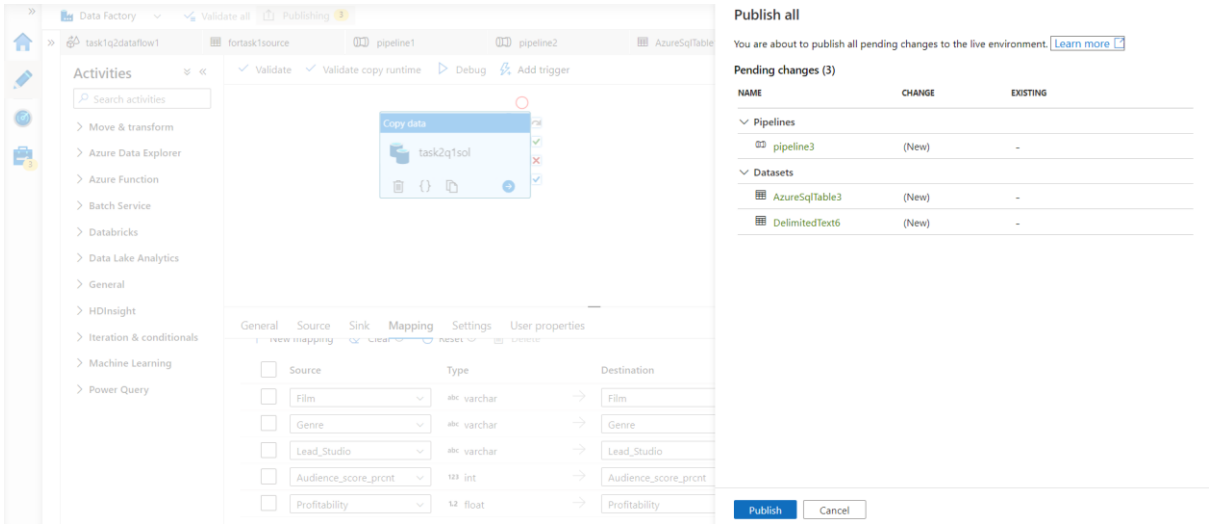


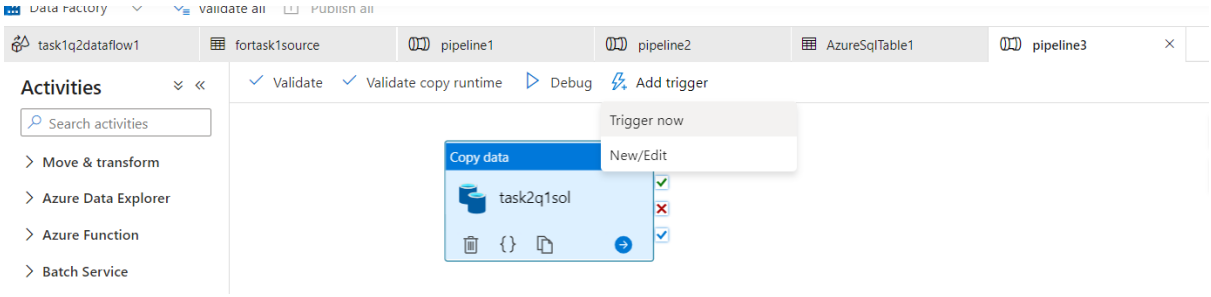

Data Factory
Validate all
Publish all
task1q2dataflow1
fortask1source
pipeline1
pipeline2
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
task2q1sol
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
Select...
New
New dataset
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more
Select a data store
Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Database for PostgreSQL
Azure Databricks Delta Lake
Azure File Storage
Azure SQL Database
Azure SQL Database Managed Instance
Azure Synapse Analytics
cassandra

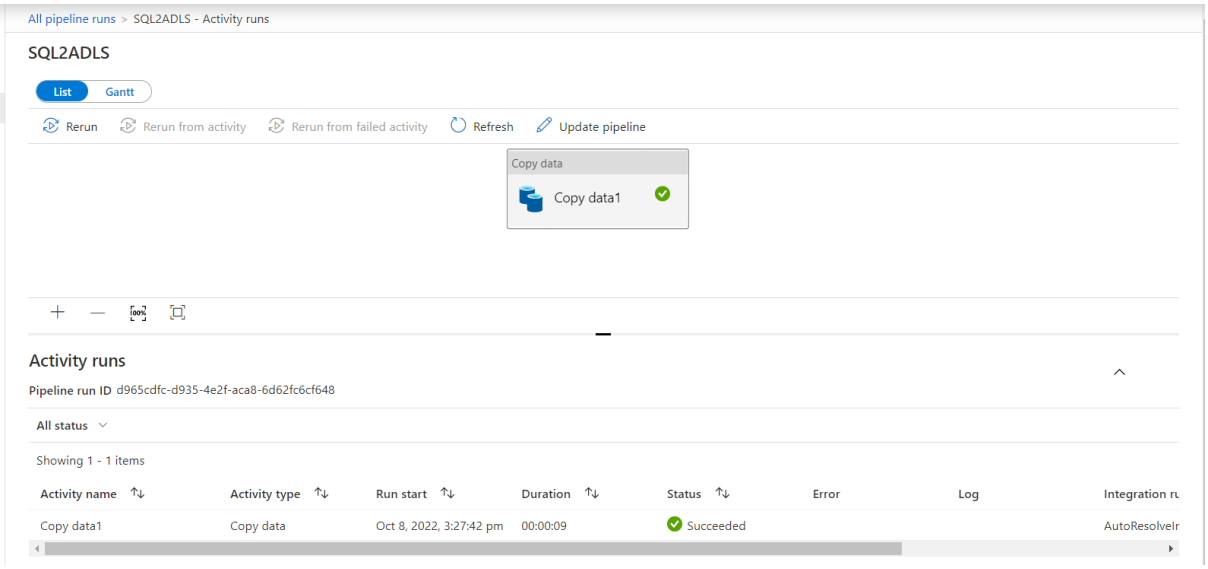

Data Factory
Validate all
Publish all
task1q2dataflow1
fortask1source
pipeline1
pipeline2
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
task2q1sol
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
Select...
New
Set properties
Name
AzureSqlTable3
Linked service *
AzureSqlDatabase1
Table name
dbo.movieoutput
Edit
Import schema
From connection/store
None
Advanced
OK
Back
Cancel

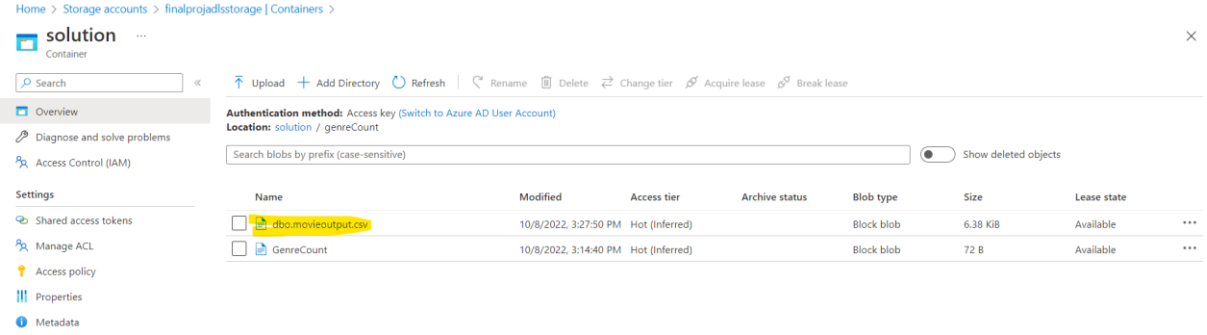

Microsoft Azure
Data Factory
vpfinalprojADF
Search
pandeyvaibhav9415@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
task1q2dataflow1
fortask1source
pipeline1
pipeline2
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
task2q1sol
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
Select...
New
Set properties
Name
DelimitedText6
Linked service *
AzureDataLakeStorage1
File path
sql2adlsfinaldest
data
File name
First row as header
Import schema
From connection/store
From sample file
None
Advanced
OK
Back
Cancel


Validate
Validate copy runtime
Debug
Add trigger
Copy data
task2q1sol
General
Source
Sink
Mapping
Settings
User properties
Film varchar Film String
Genre varchar Genre String
Lead_Studio varchar Lead_Studio String
Audience_score_prcnt int Audience_score_prcnt String
Profitability float Profitability String
Rotten_Tomatoes_prcnt int Rotten_Tomatoes_prcnt String
Worldwide_Gross varchar Worldwide_Gross String
Year int Year String
FilmCount int FilmCount String
Add dynamic content [Alt+Shift+D]
Properties
General
Related
Name *
SQL2ADLS
Description
Annotations
New


Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Publish all
You are about to publish all pending changes to the live environment.
Learn more
Pending changes (3)
NAME
CHANGE
EXISTING
Pipelines
pipeline3 (New) -
Datasets
AzureSqlTable3 (New) -
DelimitedText6 (New) -
Publish
Cancel

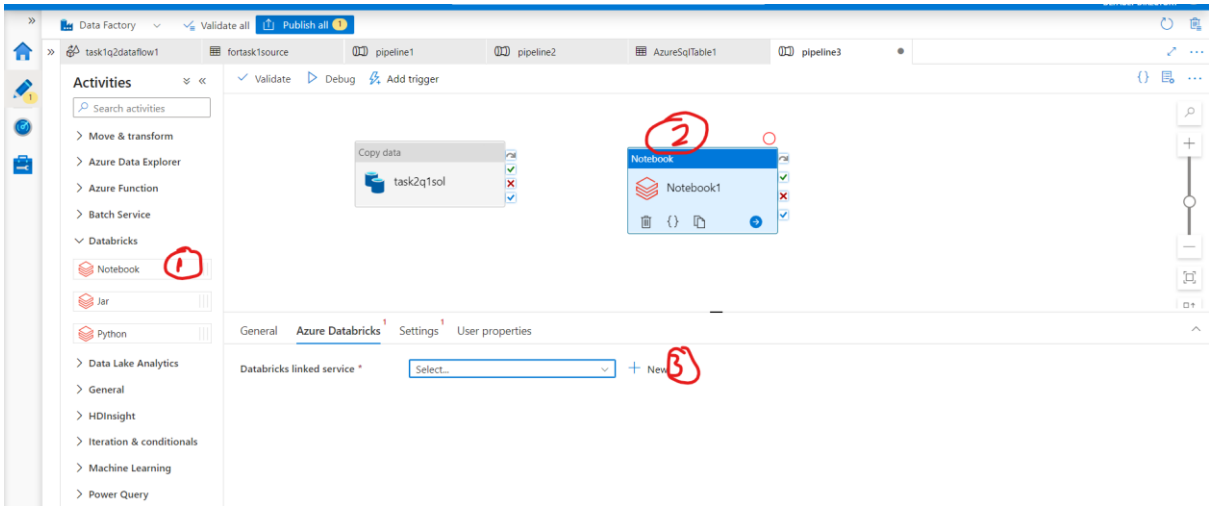

task1q2dataflow1
fortask1source
pipeline1
pipeline2
AzureSqlTable1
pipeline3
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Validate
Validate copy runtime
Debug
Add trigger
Trigger now
New/Edit
Copy data
task2q1sol

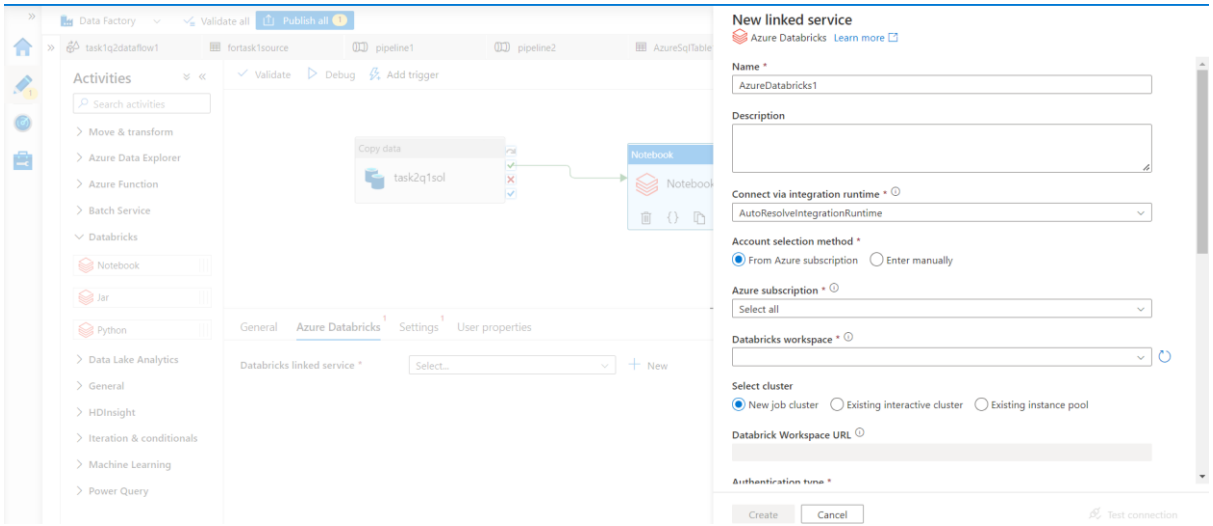


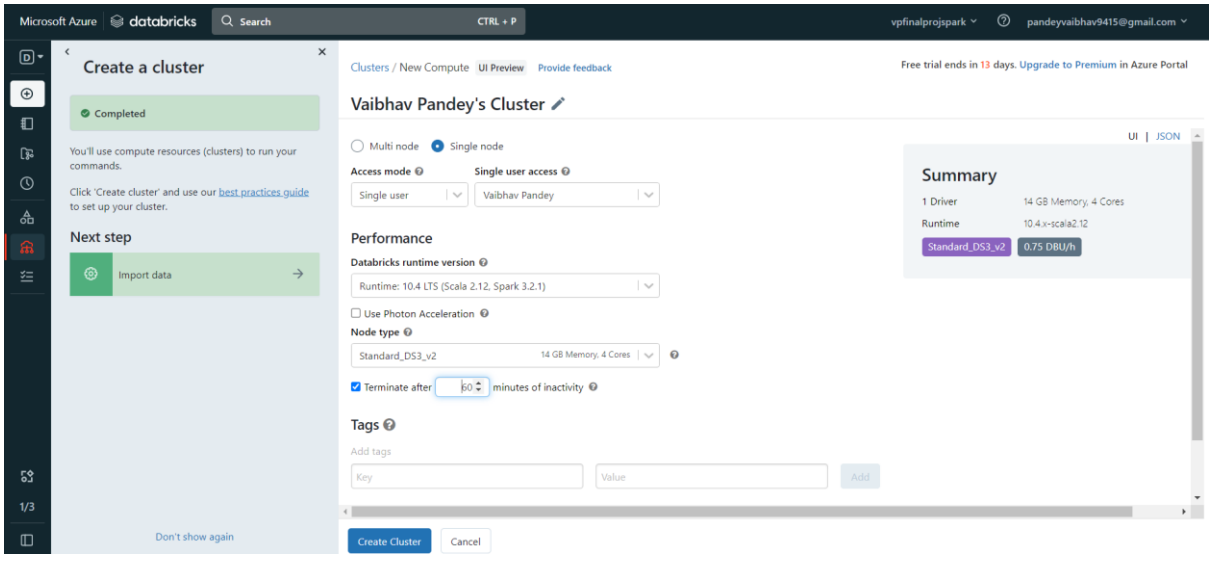


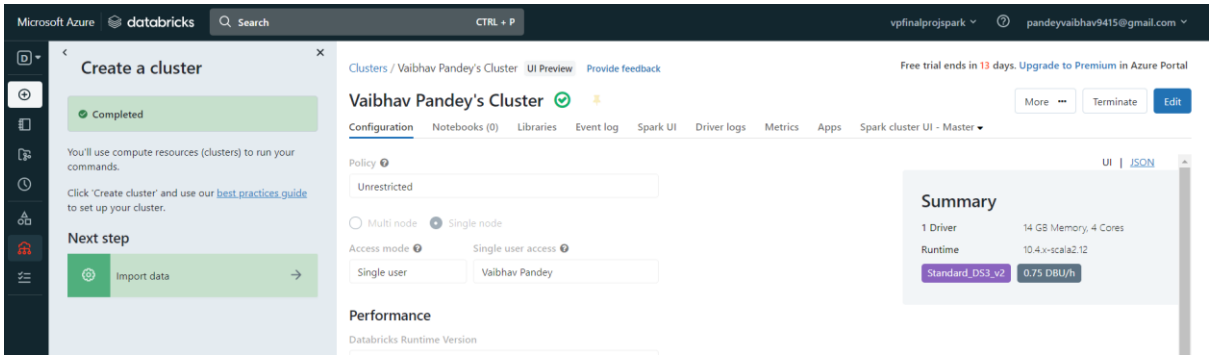
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Film	Genre	Lead_Stud	Audience_	Profitabilit	Rotten_To	Worldwide	Year	FilmCount						
2	Zack and M	Romance	The Weins	70	1.747542	64	$41.94	2008	15						
3	Waitress	Romance	Independe	67	11.08974	89	$22.18	2007	15						
4	Waiting Fo	Romance	Independe	53	0.005	6	$0.03	2011	15						
5	Tyler Perry	Romance	Independe	47	3.724192	46	$55.86	2007	15						
6	Twilight: B	Romance	Independe	68	6.383364	26	$702.17	2011	15						
7	Twilight	Romance	Summit	82	10.18003	49	$376.66	2008	15						
8	Something	Romance	Independe	48	1.719514	15	$60.18	2011	15						
9	P.S. I Love	Romance	Independe	82	5.103117	21	$153.09	2007	15						
10	One Day	Romance	Independe	54	3.682733	37	$55.24	2011	15						
11	New Year'	Romance	Warner Br	48	2.536429	8	$142.04	2011	15						
12	Music and	Romance	Warner Br	70	3.647411	63	$145.90	2007	15						
13	Monte Car	Romance	20th Centu	50	1.9832	38	$39.66	2011	15						
14	Midnight ir	Romance	Sony	84	8.744706	93	$148.66	2011	15						
15	Jane Eyre	Romance	Universal	77	0	85	$30.15	2011	15						
16	Across the	Romance	Independe	84	0.652603	54	$29.37	2007	15						
17	The Curiou	Fantasy	Warner Br	81	1.783944	73	$285.43	2008	1						
18	Water For	Drama	20th Centu	72	3.081421	60	$117.09	2011	13						
19	The Twilig	Drama	Summit	78	14.1964	27	$709.82	2009	13						
20	The Time T	Drama	Paramoun	65	2.598205	38	$101.33	2009	13						
21	The Duche	Drama	Paramoun	68	3.20785	60	$43.31	2008	13						
22	Remembe	Drama	Summit	70	3.49125	28	$55.86	2010	13						
23	Rachel Get	Drama	Independe	61	1.384167	85	$16.61	2008	13						
24	Not Easily	Drama	Independe	66	2.14	34	$10.70	2009	13						
25	My Week	Drama	The Weins	84	0.8258	83	$8.26	2011	13						
26	Love Happ	Drama	Universal	40	2.004444	18	$36.08	2009	13						

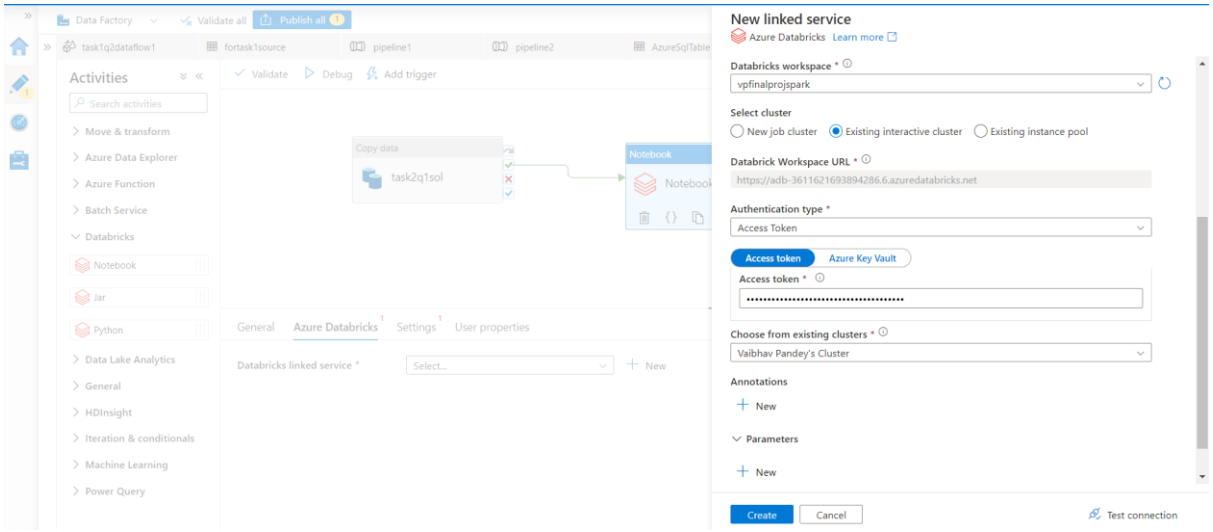
2. Create an activity that can use Azure Databricks to read the data from the ADLS Gen2 and perform rank operation on the Genre column. Ensure this activity gets activated only after the data is stored in ADLS from SQL DB. The result of Databricks must be stored in the ADLS.

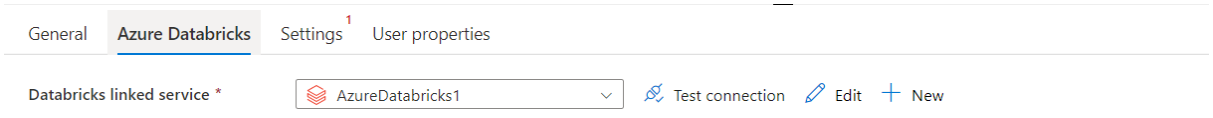


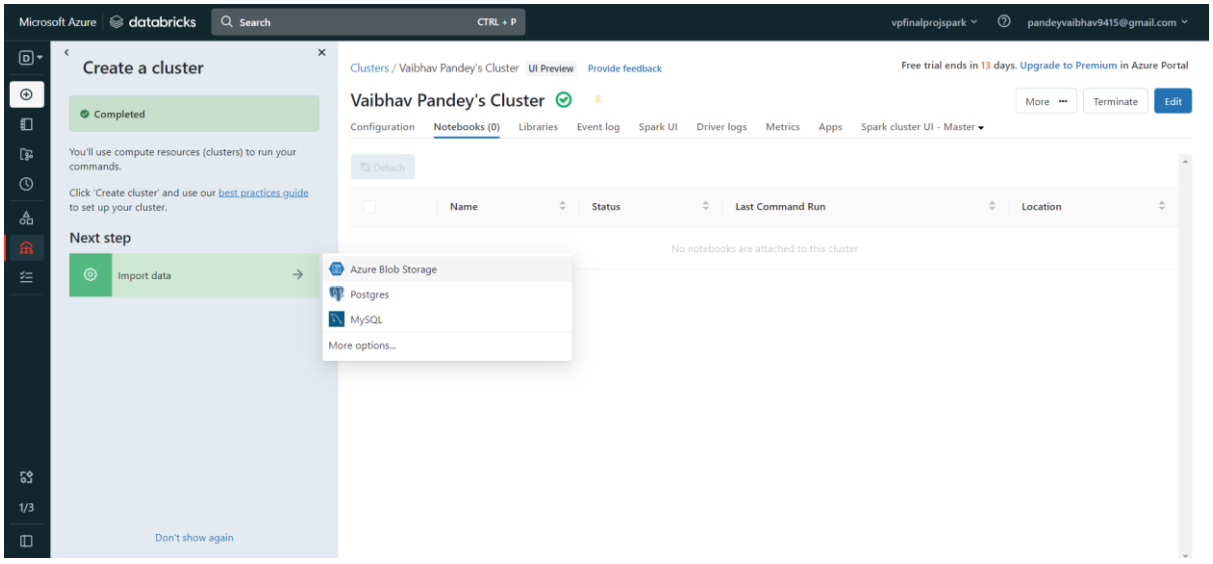


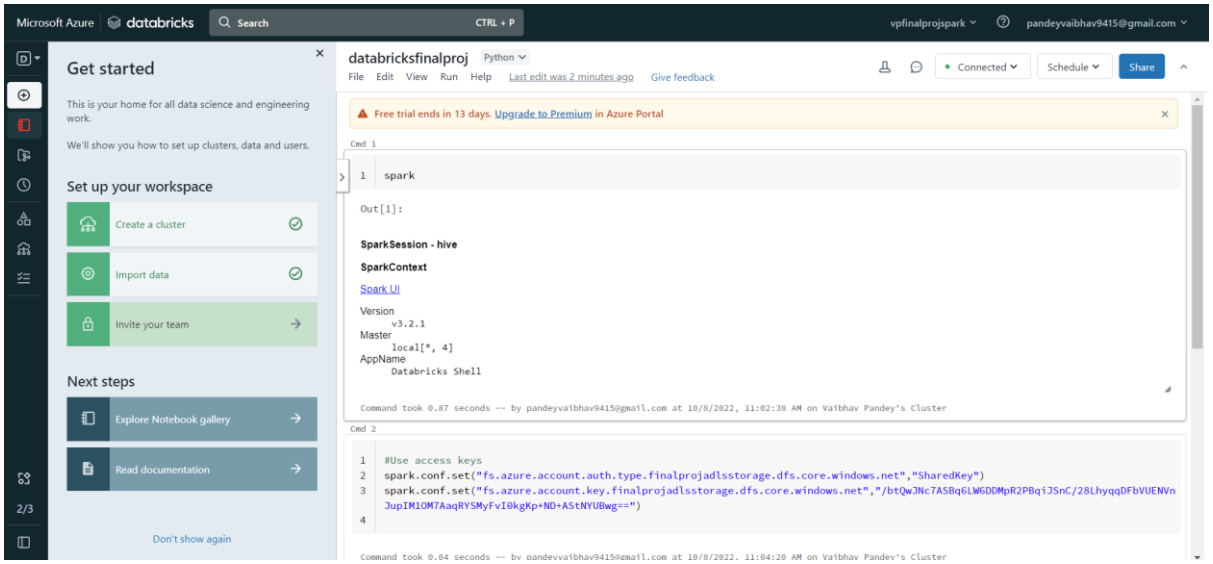


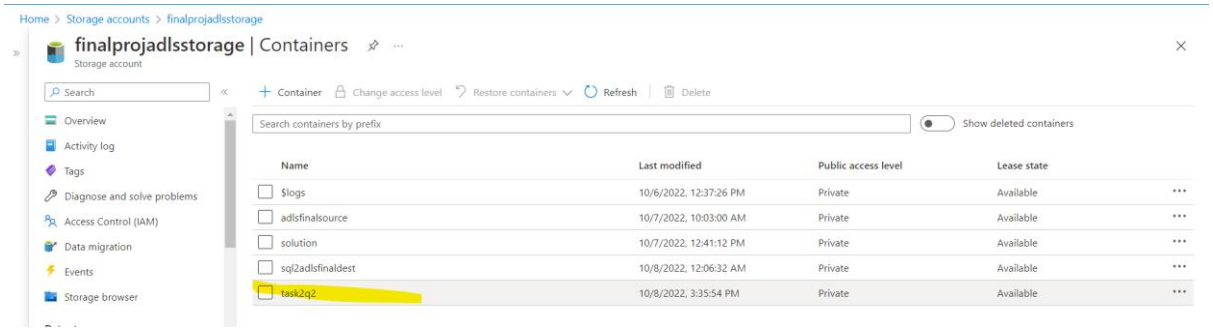

Microsoft Azure
databricks
Search
CTRL + P
vpfinalprojspark
pandeyvaibhav9415@gmail.com
Create a cluster
Completed
You'll use compute resources (clusters) to run your commands.
Click 'Create cluster' and use our best practices guide to set up your cluster.
Next step
Import data
Clusters / Vaibhav Pandey's Cluster
UI Preview
Provide feedback
Free trial ends in 13 days. Upgrade to Premium in Azure Portal
Vaibhav Pandey's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark cluster UI - Master
Policy
Unrestricted
Multi node
Single node
Access mode
Single user access
Single user
Vaibhav Pandey
Performance
Databricks Runtime Version
UI | JSON
Summary
1 Driver
14 GB Memory, 4 Cores
Runtime
10.4.x-scala2.12
Standard_DS3_v2
0.75 DBU/h

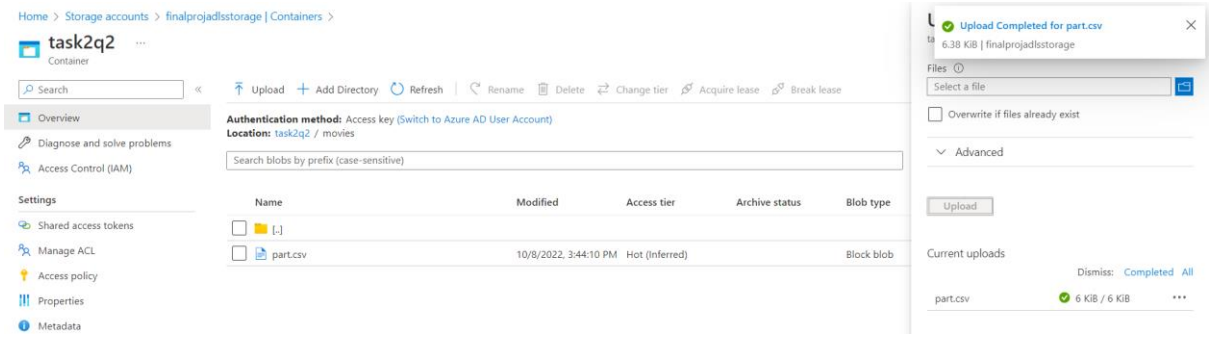

Data Factory
Validate all
Publish all
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Databricks
Notebook
Jar
Python
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
task2q1sol
Notebook
General
Azure Databricks
Settings
User properties
Databricks linked service *
Select...
New
New linked service
Azure Databricks
Learn more
Databricks workspace *
vpfinalprojspark
Select cluster
New job cluster
Existing interactive cluster
Existing instance pool
Databrick Workspace URL *
https://adb-3611621693894286.6.azuredatabricks.net
Authentication type *
Access Token
Access token
Azure Key Vault
Access token *
Choose from existing clusters *
Vaibhav Pandey's Cluster
Annotations
New
Parameters
New
Create
Cancel
Test connection

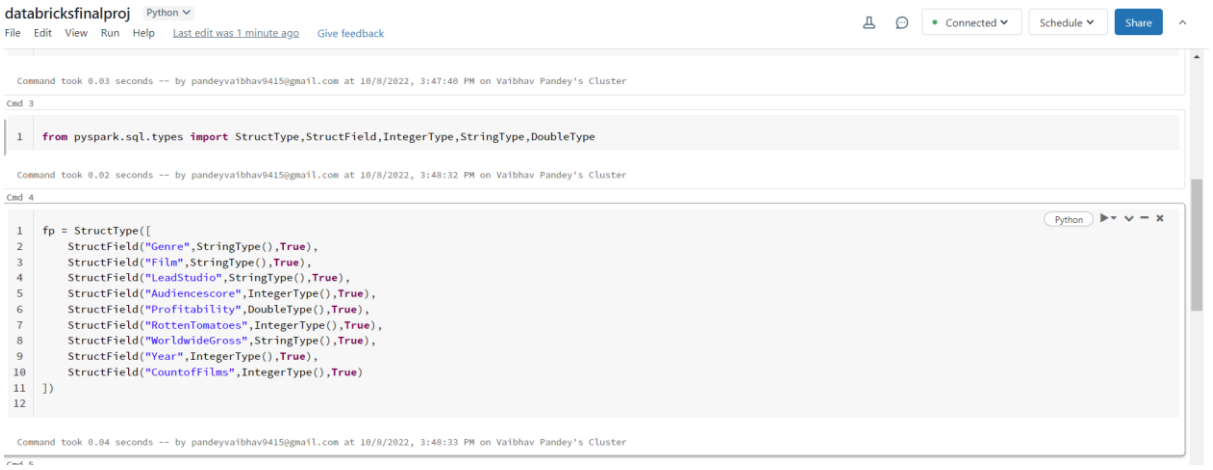

General
Azure Databricks
Settings
1
User properties
Databricks linked service *
AzureDatabricks1
Test connection
Edit
New


Microsoft Azure
databricks
Search
CTRL + P
vpfinalprojspark
pandeyvaibhav9415@gmail.com
Create a cluster
Completed
You'll use compute resources (clusters) to run your commands.
Click 'Create cluster' and use our best practices guide to set up your cluster.
Next step
Import data
Azure Blob Storage
Postgres
MySQL
More options...
Clusters / Vaibhav Pandey's Cluster
UI Preview
Provide feedback
Free trial ends in 13 days. Upgrade to Premium in Azure Portal
Vaibhav Pandey's Cluster
More
Terminate
Edit
Configuration
Notebooks (0)
Libraries
Event log
Spark UI
Driver logs
Metrics
Apps
Spark cluster UI - Master
Detach
Name
Status
Last Command Run
Location
No notebooks are attached to this cluster
1/3
Don't show again


Microsoft Azure
databricks
Search
CTRL + P
vpfinalprojspark
pandeyvaibhav9415@gmail.com
Get started
This is your home for all data science and engineering work.
We'll show you how to set up clusters, data and users.
Set up your workspace
Create a cluster
Import data
Invite your team
Next steps
Explore Notebook gallery
Read documentation
Don't show again
databricksfinalproj
Python
File Edit View Run Help
Last edit was 2 minutes ago
Give feedback
Connected
Schedule
Share
Free trial ends in 13 days. Upgrade to Premium in Azure Portal
Cmd 1
1 spark
Out[1]:
SparkSession - hive
SparkContext
Spark UI
Version
v3.2.1
Master
local[*, 4]
AppName
Databricks Shell
Command took 0.87 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 11:02:38 AM on Vaibhav Pandey's Cluster
Cmd 2
1 #Use access keys
2 spark.conf.set("fs.azure.account.auth.type.finalprojadlsstorage.dfs.core.windows.net","SharedKey")
3 spark.conf.set("fs.azure.account.key.finalprojadlsstorage.dfs.core.windows.net","/btQwJNc7ASBq6LW6DDMpR2PBqiJSnC/28LhyqqDFbVUENVnJupIM1OM7AaqRYSMyFvI0kgKp+ND+AStNYUBwg==")
4
Command took 0.04 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 11:04:20 AM on Vaibhav Pandey's Cluster


Home > Storage accounts > finalprojadlsstorage
finalprojadlsstorage | Containers
Storage account
Search
Container
Change access level
Restore containers
Refresh
Delete
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Search containers by prefix
Show deleted containers
Name
Last modified
Public access level
Lease state
$logs
10/6/2022, 12:37:26 PM
Private
Available
adlsfinalsource
10/7/2022, 10:03:00 AM
Private
Available
solution
10/7/2022, 12:41:12 PM
Private
Available
sql2adlsfinaldest
10/8/2022, 12:06:32 AM
Private
Available
task2q2
10/8/2022, 3:35:54 PM
Private
Available


Home > Storage accounts > finalprojadlsstorage | Containers >
task2q2
Container
Search
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Authentication method: Access key (Switch to Azure AD User Account)
Location: task2q2 / movies
Search blobs by prefix (case-sensitive)
Name
Modified
Access tier
Archive status
Blob type
[..]
part.csv
10/8/2022, 3:44:10 PM
Hot (Inferred)
Block blob
Upload Completed for part.csv
6.38 KiB | finalprojadlsstorage
Files
Select a file
Overwrite if files already exist
Advanced
Upload
Current uploads
Dismiss: Completed All
part.csv
6 KiB / 6 KiB


databricksfinalproj
Python
File Edit View Run Help
Last edit was 1 minute ago
Give feedback
Connected
Schedule
Share
Command took 0.03 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 3:47:40 PM on Vaibhav Pandey's Cluster
Cmd 3
1 from pyspark.sql.types import StructType,StructField,IntegerType,StringType,DoubleType
Command took 0.02 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 3:48:32 PM on Vaibhav Pandey's Cluster
Cmd 4
Python
1 fp = StructType([
2 StructField("Genre",StringType(),True),
3 StructField("Film",StringType(),True),
4 StructField("LeadStudio",StringType(),True),
5 StructField("Audiencescore",IntegerType(),True),
6 StructField("Profitability",DoubleType(),True),
7 StructField("RottenTomatoes",IntegerType(),True),
8 StructField("WorldwideGross",StringType(),True),
9 StructField("Year",IntegerType(),True),
10 StructField("CountofFilms",IntegerType(),True)
11])
12
Command took 0.04 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 3:48:33 PM on Vaibhav Pandey's Cluster

```
df=spark.read.option('header',True).option('inferSchema',True).csv("abfss://task2q2@finalprojadlsstorage.dfs.core.windows.net/movies/part.csv")
df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [Film: string, Genre: string ... 7 more fields]

Film	Genre	Lead_Studio	Audience_score_prcnt	Profitability	Rotten_Tomatoes_prcnt	Worldwide_Gross	Year	FilmCount
Zack and Miri Mak...	Romance	The Weinstein Com...	70	1.747541667	64	$41.94	2008	15
Waitress	Romance	Independent	67	11.0897415	89	$22.18	2007	15
Waiting For Forever	Romance	Independent	53	0.005	6	$0.03	2011	15
Tyler Perry's Why...	Romance	Independent	47	3.7241924	46	$55.86	2007	15
Twilight: Breakin...	Romance	Independent	68	6.383363636	26	$702.17	2011	15
Twilight	Romance	Summit	82	10.18002703	49	$376.66	2008	15
Something Borrowed	Romance	Independent	48	1.719514286	15	$60.18	2011	15
P.S. I Love You	Romance	Independent	82	5.103116833	21	$153.09	2007	15
One Day	Romance	Independent	54	3.682733333	37	$55.24	2011	15
New Year's Eve	Romance	Warner Bros.	48	2.536428571	8	$142.04	2011	15
Music and Lyrics	Romance	Warner Bros.	70	3.64741055	63	$145.90	2007	15
Monte Carlo	Romance	20th Century Fox	50	1.9832	38	$39.66	2011	15
Midnight in Paris	Romance	Sony	84	8.744705882	93	$148.66	2011	15
Jane Eyre	Romance	Universal	77	0.0	85	$30.15	2011	15
Across the Universe	Romance	Independent	84	0.652603178	54	$29.37	2007	15
The Curious Case ...	Fantasy	Warner Bros.	81	1.78394375	73	$285.43	2008	1
Water For Elephants	Drama	20th Century Fox	72	3.081421053	60	$117.09	2011	13
The Twilight Saga...	Drama	Summit	78	14.1964	27	$709.82	2009	13

Command took 4.51 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 3:48:35 PM on Vaibhav Pandey's Cluster

```
df.registerTempTable("movies")
```

```
/databricks/spark/python/pyspark/sql/dataframe.py:146: FutureWarning: Deprecated in 2.0, use createOrReplaceTempView instead.
  warnings.warn(
```

Command took 0.12 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 3:49:54 PM on Vaibhav Pandey's Cluster

Cmd 7

Python

```
op= spark.sql("select *,RANK () OVER(ORDER BY Genre) as Rankedby_Genre from movies")
op.show()
```

(1) Spark Jobs

op: pyspark.sql.dataframe.DataFrame = [Film: string, Genre: string ... 8 more fields]

Film	Genre	Lead_Studio	Audience_score_prcnt	Profitability	Rotten_Tomatoes_prcnt	Worldwide_Gross	Year	FilmCount	Rankedby_Genre
Killers	Action	Lionsgate	45	1.245333333	11	$93.40	2010	1	1
WALL-E	Animation	Disney	89	2.896019067	96	$521.28	2008	4	2
Tangled	Animation	Disney	88	1.365692308	89	$355.01	2010	4	2
Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	$193.97	2011	4	2
Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	$193.97	2011	4	2
Sex and the City Two	Comedy	Warner Bros.	49	2.8835	15	$288.35	2010	43	6
The Heartbreak Kid	Comedy	Paramount	41	2.129444167	30	$127.77	2007	43	6
Sex and the City 2	Comedy	Warner Bros.	49	2.8835	15	$288.35	2010	43	6
Youth in Revolt	Comedy	The Weinstein Com...	52	1.09	68	$19.62	2010	43	6
Sex and the City	Comedy	Warner Bros.	81	7.221795791	49	$415.25	2008	43	6
When in Rome	Comedy	Disney	44	0.0	15	$43.04	2010	43	6
Penelope	Comedy	Summit	74	1.382799733	52	$20.74	2008	43	6
Valentine's Day	Comedy	Warner Bros.	54	4.184038462	17	$217.57	2010	43	6
Over Her Dead Body	Comedy	New Line	47	2.071	15	$20.71	2008	43	6
The Proposal	Comedy	Disney	74	7.8675	43	$314.70	2009	43	6
Our Family Wedding	Comedy	Independent	49	0.0	14	$21.37	2010	43	6
Miss Pettigrew Li...	Comedy	Independent	70	0.2528949	78	$15.17	2008	43	6
Marley and Me	Comedy	Fox	77	3.746781818	63	$206.07	2008	43	6

SQL

```
%sql
select *,RANK () OVER(ORDER BY Genre) as Rankedby_Genre from movies;
```

(1) Spark Jobs

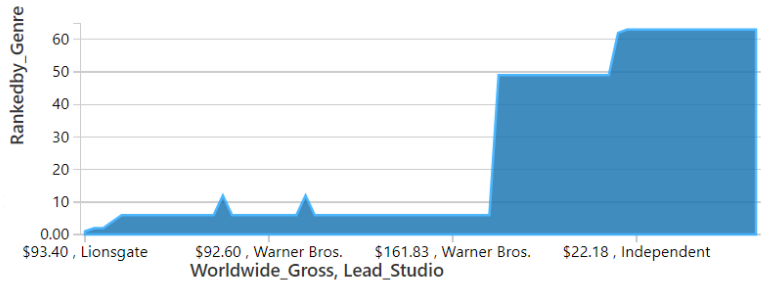
_sqldf: pyspark.sql.dataframe.DataFrame = [Film: string, Genre: string ... 8 more fields]

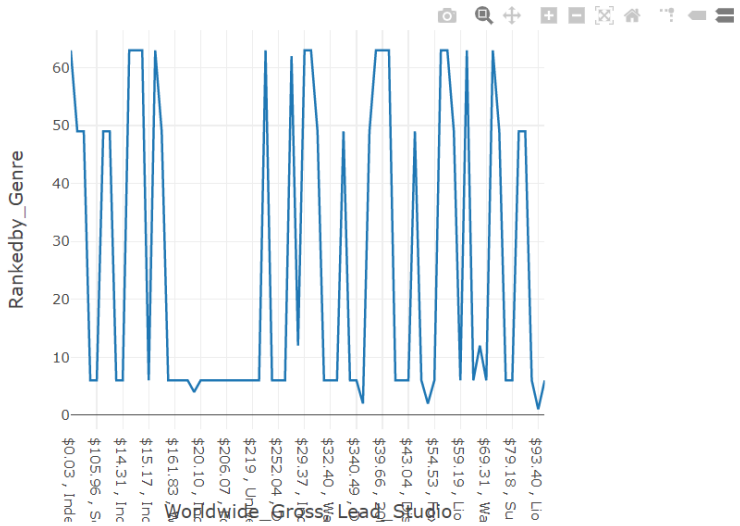
Table | Data Profile

	Film	Genre	Lead_Studio	Audience_score_prcnt	Profitability	Rotten_Tomatoes_prcnt	Worldwide_Gross	Year	FilmCount
1	Killers	Action	Lionsgate	45	1.245333333	11	$93.40	2010	1
2	WALL-E	Animation	Disney	89	2.896019067	96	$521.28	2008	4
3	Tangled	Animation	Disney	88	1.365692308	89	$355.01	2010	4
4	Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	$193.97	2011	4
5	Gnomeo and Juliet	Animation	Disney	52	5.387972222	56	$193.97	2011	4
6	Youth in Revolt	Comedy	The Weinstein Company	52	1.09	68	$19.62	2010	43
7	You Will Meet a Tall Dark Stranger	Comedy	Independent	35	1.211818182	43	$26.66	2010	43

Showing all 77 rows.

Chart | Data Profile





	Rotten_Tomatoes_prcnt, Worldwide_Gross	FilmCount	Rankedby_Genre
1	11, $93.40	1	1
2	96, $521.28	4	2
3	89, $355.01	4	2
4	56, $193.97	8	4
5	68, $19.62	43	6
6	43, $26.66	43	6
7	15, $43.04	43	6

Plot Options...

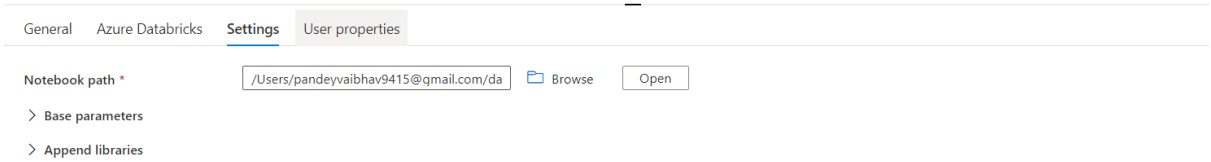
Validate | Debug | Add trigger

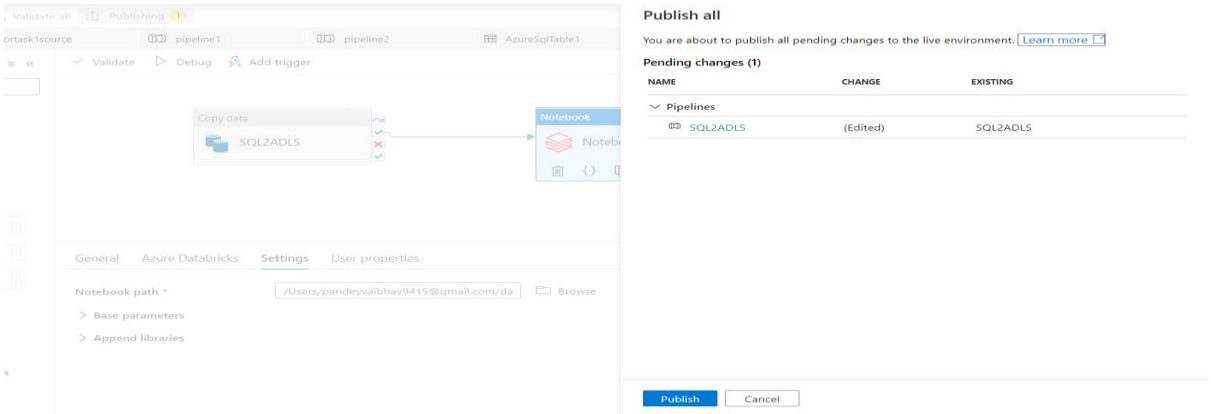
General | Azure Databricks | Settings | User properties

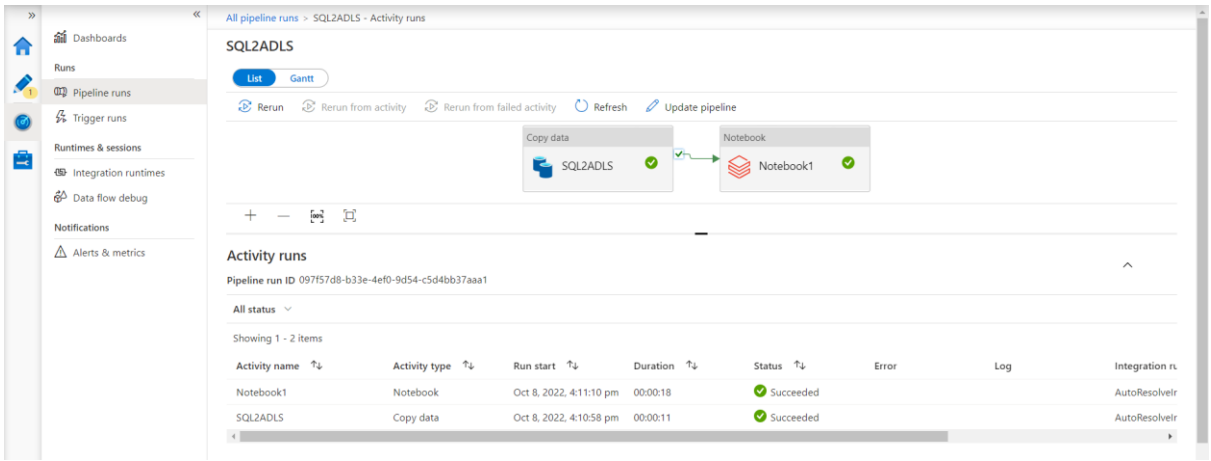
Databricks linked service * AzureDatabricks1 Test connection Edit New

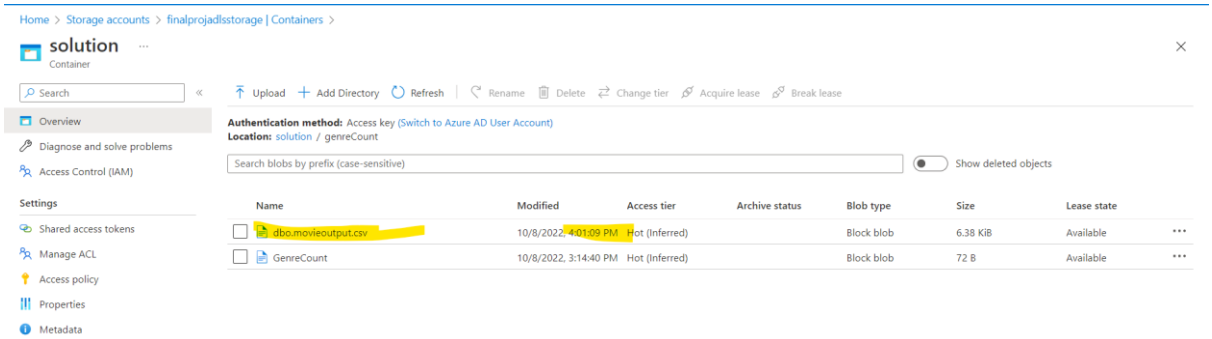
Connection successful


Copy data
SQL2ADLS
Notebook
Notebook1

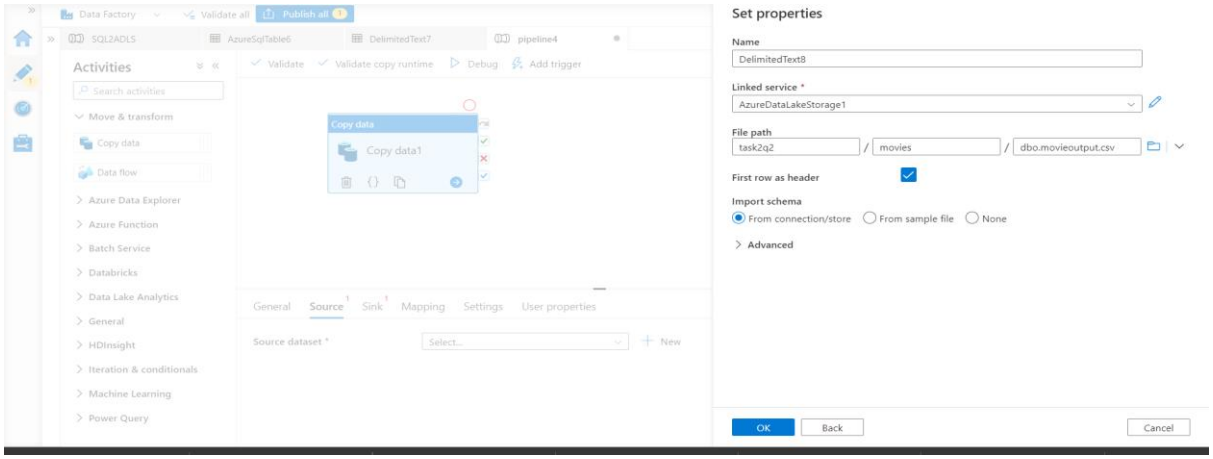

General
Azure Databricks
Settings
User properties
Notebook path *
/Users/pandeyvaibhav9415@gmail.com/da
Browse
Open
Base parameters
Append libraries

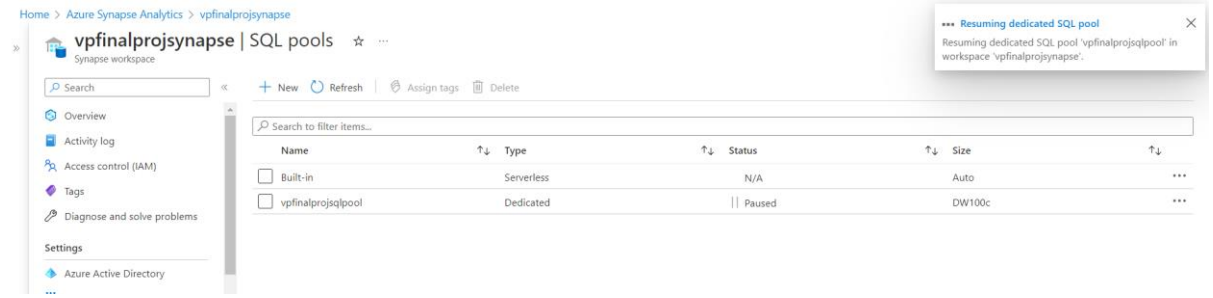

Publish all
You are about to publish all pending changes to the live environment.
Learn more
Pending changes (1)
NAME
CHANGE
EXISTING
Pipelines
SQL2ADLS
(Edited)
SQL2ADLS
Publish
Cancel

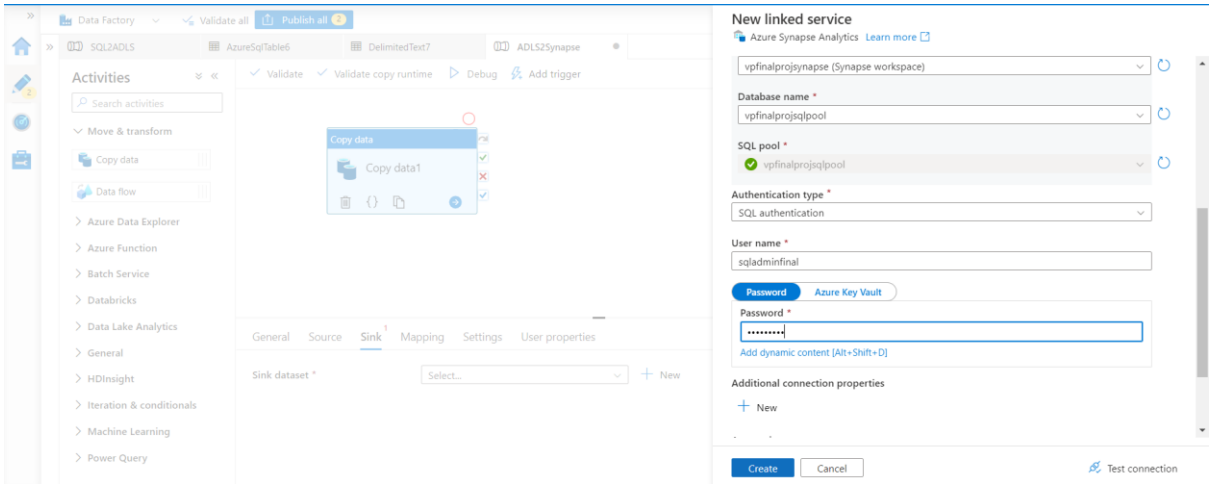

Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics
All pipeline runs > SQL2ADLS - Activity runs
SQL2ADLS
List
Gantt
Rerun
Rerun from activity
Rerun from failed activity
Refresh
Update pipeline
Copy data
SQL2ADLS
Notebook
Notebook1
Activity runs
Pipeline run ID 097f57d8-b33e-4ef0-9d54-c5d4bb37aaa1
All status
Showing 1 - 2 items
Activity name
Activity type
Run start
Duration
Status
Error
Log
Integration ru
Notebook1
Notebook
Oct 8, 2022, 4:11:10 pm
00:00:18
Succeeded
AutoResolveIr
SQL2ADLS
Copy data
Oct 8, 2022, 4:10:58 pm
00:00:11
Succeeded
AutoResolveIr

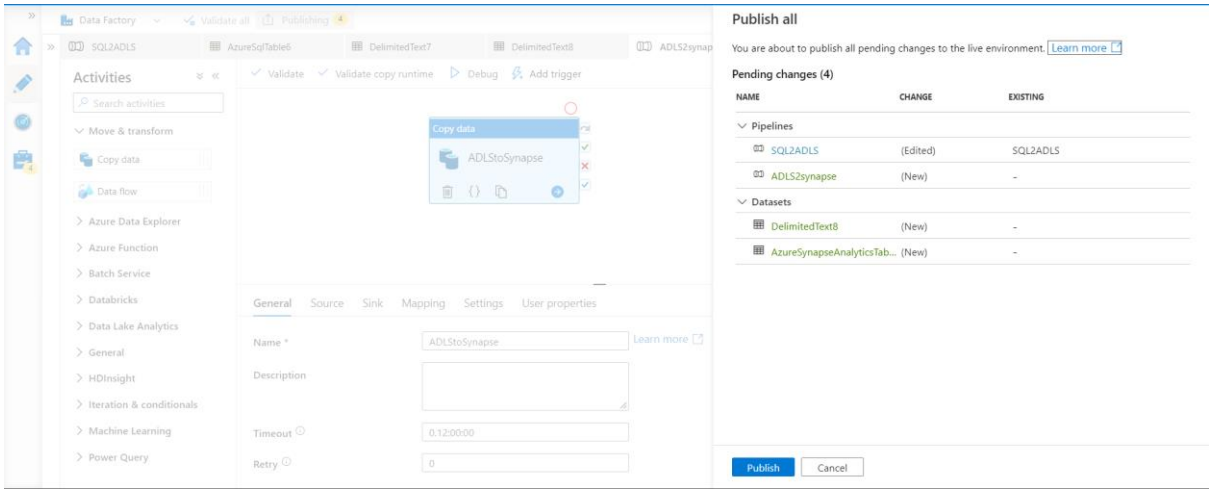

Home > Storage accounts > finalprojadlsstorage | Containers >
solution
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Authentication method: Access key (Switch to Azure AD User Account)
Location: solution / genreCount
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
dbo.movieoutput.csv
10/8/2022, 4:01:09 PM
Hot (Inferred)
Block blob
6.38 KiB
Available
GenreCount
10/8/2022, 3:14:40 PM
Hot (Inferred)
Block blob
72 B
Available

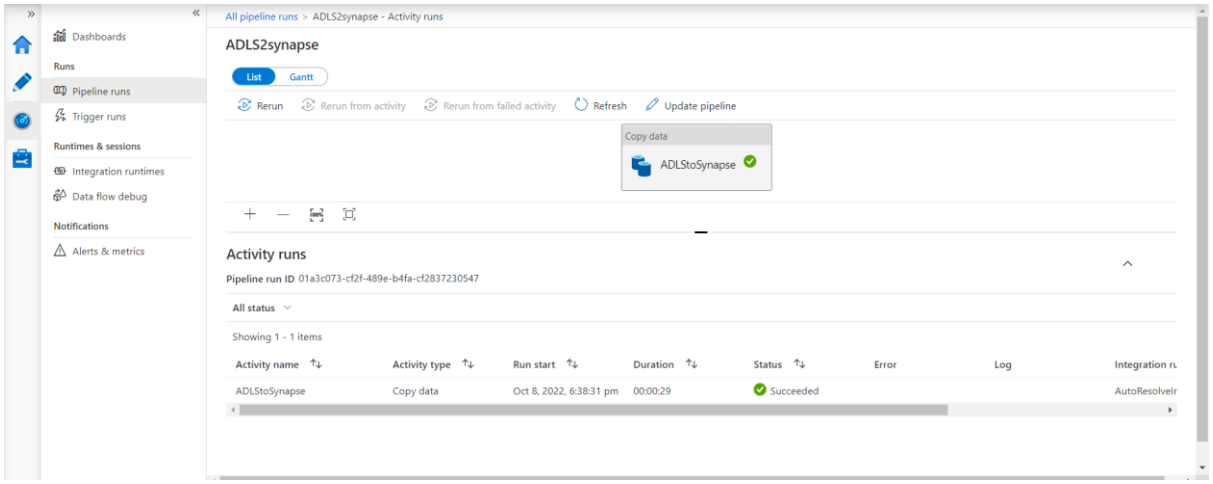
3. Create a final activity that will read the output of previous activity in ADLS and store the same in Synapse.

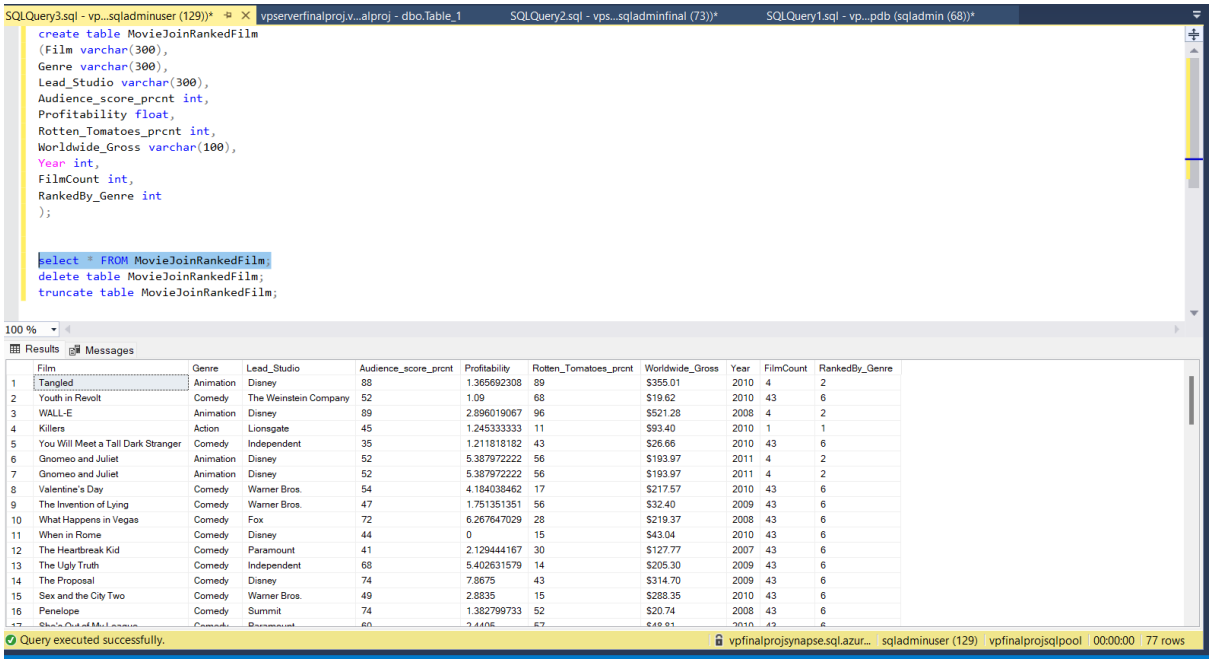







Publish all
You are about to publish all pending changes to the live environment. Learn more
Pending changes (4)
NAME
CHANGE
EXISTING
Pipelines
SQL2ADLS
(Edited)
SQL2ADLS
ADLS2synapse
(New)
-
Datasets
DelimitedText8
(New)
-
AzureSynapseAnalyticsTab...
(New)
-
Publish
Cancel
Activities
Copy data
ADLStoSynapse
General
Source
Sink
Mapping
Settings
User properties
Name *
Description
Timeout
Retry


All pipeline runs > ADLS2synapse - Activity runs
ADLS2synapse
List
Gantt
Rerun
Rerun from activity
Rerun from failed activity
Refresh
Update pipeline
Copy data
ADLStoSynapse
Activity runs
Pipeline run ID 01a3c073-cf2f-489e-b4fa-cf2837230547
All status
Showing 1 - 1 items
Activity name
Activity type
Run start
Duration
Status
Error
Log
ADLStoSynapse
Copy data
Oct 8, 2022, 6:38:31 pm
00:00:29
Succeeded
Dashboards
Runs
Pipeline runs
Trigger runs
Runtimes & sessions
Integration runtimes
Data flow debug
Notifications
Alerts & metrics


SQLQuery3.sql - vp...sqladminuser (129))*
vpserverfinalproj.v...alproj - dbo.Table_1
SQLQuery2.sql - vps...sqladminfinal (73))*
SQLQuery1.sql - vp...pdb (sqladmin (68))*
create table MovieJoinRankedFilm
(Film varchar(300),
Genre varchar(300),
Lead_Studio varchar(300),
Audience_score_prcnt int,
Profitability float,
Rotten_Tomatoes_prcnt int,
Worldwide_Gross varchar(100),
Year int,
FilmCount int,
RankedBy_Genre int
);
select * FROM MovieJoinRankedFilm;
delete table MovieJoinRankedFilm;
truncate table MovieJoinRankedFilm;
100 %
Results
Messages
Film
Genre
Lead_Studio
Audience_score_prcnt
Profitability
Rotten_Tomatoes_prcnt
Worldwide_Gross
Year
FilmCount
RankedBy_Genre
1 Tangled Animation Disney 88 1.365692308 89 $355.01 2010 4 2
2 Youth in Revolt Comedy The Weinstein Company 52 1.09 68 $19.62 2010 43 6
3 WALL-E Animation Disney 89 2.896019067 96 $521.28 2008 4 2
4 Killers Action Lionsgate 45 1.245333333 11 $93.40 2010 1 1
5 You Will Meet a Tall Dark Stranger Comedy Independent 35 1.211818182 43 $26.66 2010 43 6
6 Gnomeo and Juliet Animation Disney 52 5.387972222 56 $193.97 2011 4 2
7 Gnomeo and Juliet Animation Disney 52 5.387972222 56 $193.97 2011 4 2
8 Valentine's Day Comedy Warner Bros. 54 4.184038462 17 $217.57 2010 43 6
9 The Invention of Lying Comedy Warner Bros. 47 1.751351351 56 $32.40 2009 43 6
10 What Happens in Vegas Comedy Fox 72 6.267647029 28 $219.37 2008 43 6
11 When in Rome Comedy Disney 44 0 15 $43.04 2010 43 6
12 The Heartbreak Kid Comedy Paramount 41 2.129444167 30 $127.77 2007 43 6
13 The Ugly Truth Comedy Independent 68 5.402631579 14 $205.30 2009 43 6
14 The Proposal Comedy Disney 74 7.8675 43 $314.70 2009 43 6
15 Sex and the City Two Comedy Warner Bros. 49 2.8835 15 $288.35 2010 43 6
16 Penelope Comedy Summit 74 1.382799733 52 $20.74 2008 43 6
Query executed successfully.
vpfinalprojsynapse.sql.azur...
sqladminuser (129)
vpfinalprojsqlpool
00:00:00
77 rows