

# 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.

The screenshot shows the 'Create a storage account' wizard in the Azure portal. The 'Basics' tab is selected. The page title is 'Create a storage account'. Below the title, there's a brief description of Azure Storage. The 'Project details' section asks for a subscription and a resource group. The subscription dropdown shows 'Azure Pass - Sponsorship'. The resource group dropdown shows '(New) vp06102022' with a 'Create new' link below it.

## ● Create a Storage account.

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

The screenshot shows the Microsoft Azure home page. At the top, there's a navigation bar with various icons and links. Below it, the main content area has a section titled "Azure services" with a grid of icons. One icon, "Storage accounts", is highlighted with a red box. Other icons include "SQL databases", "Azure Synapse Analytics", "Data factories", "Subscriptions", "Azure Databricks", "Quickstart Center", "Virtual machines", and "More services". Below this is a "Resources" section with "Recent" and "Favorite" tabs.

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

The screenshot shows the "Storage accounts" list page. At the top, there's a header with "Storage accounts" and a "Create" button. Below the header are filter options and grouping controls. The main area displays a list of storage accounts, with one entry visible: "(New) vp06102022".

Select redundancy as 'LRS' and Performance as Standard

The screenshot shows the "Create a storage account" wizard on the "Basics" step. The top navigation bar includes "Review" and "Next : Advanced >". The "Basics" tab is selected. Under "Instance details", there are fields for "Storage account name" (set to "finalproj1storage") and "Region" (set to "(US) East US"). For "Performance", the "Standard" radio button is selected. For "Redundancy", the "Locally-redundant storage (LRS)" option is chosen. At the bottom, there are "Review" and "Next : Advanced" buttons.

Home > Storage accounts >

## Create a storage account ...

Basics   Advanced   Networking   Data protection   Encryption   Tags   Review

### Security

Configure security settings that impact your storage account.

|                                                                       |                                     |
|-----------------------------------------------------------------------|-------------------------------------|
| Require secure transfer for REST API operations ⓘ                     | <input checked="" type="checkbox"/> |
| Allow enabling public access on containers ⓘ                          | <input checked="" type="checkbox"/> |
| Enable storage account key access ⓘ                                   | <input checked="" type="checkbox"/> |
| Default to Azure Active Directory authorization in the Azure portal ⓘ | <input type="checkbox"/>            |
| Minimum TLS version ⓘ                                                 | Version 1.2                         |
| Permitted scope for copy operations (preview) ⓘ                       | From any storage account            |

Review

< Previous

Next : Networking >

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

Home > Storage accounts >

## Create a storage account ...

Basics   Advanced   Networking   Data protection   Encryption   Tags   Review

### Data Lake Storage Gen2

The Data Lake Storage Gen2 hierarchical namespace accelerates big data analytics workloads and enables file-level access control lists (ACLs). [Learn more](#)

Enable hierarchical namespace

### Blob storage

|                                  |                                                                                                                                                                                                                                               |
|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Enable SFTP ⓘ                    | <input type="checkbox"/>                                                                                                                                                                                                                      |
| Enable network file system v3 ⓘ  | <input type="checkbox"/>                                                                                                                                                                                                                      |
| Allow cross-tenant replication ⓘ | <input type="checkbox"/><br><small> ⓘ Cross-tenant replication and hierarchical namespace cannot be enabled simultaneously.</small>                                                                                                           |
| Access tier ⓘ                    | <input checked="" type="radio"/> Hot: Frequently accessed data and day-to-day usage scenarios<br><input type="radio"/> Cool: Infrequently accessed data and long-term storage needs<br><input type="radio"/> Archive: Long-term storage needs |

## Create a storage account ...

Basics Advanced Networking Data protection Encryption Tags Review

### Network access \*

Enable public access from all networks

Enable public access from selected virtual networks and IP addresses

Disable public access and use private access

Enabling public access from all networks might make this resource available publicly. Unless public access is required, we recommend using a more restricted access type. [Learn more](#)

### Network routing

Determine how to route your traffic as it travels from the source to its Azure endpoint. Microsoft network routing is recommended for most customers.

Routing preference ⓘ \*

Microsoft network routing

Internet routing

**Review**

< Previous

Next : Data protection >

## Create a storage account ...

Basics Advanced Networking Data protection Encryption Tags Review

### Recovery

Protect your data from accidental or erroneous deletion or modification.

Enable point-in-time restore for containers

Use point-in-time restore to restore one or more containers to an earlier state. If point-in-time restore is enabled, then versioning, change feed, and blob soft delete must also be enabled. [Learn more](#)

Enable soft delete for blobs

Soft delete enables you to recover blobs that were previously marked for deletion, including blobs that were overwritten. [Learn more](#)

Days to retain deleted blobs ⓘ

7

Enable soft delete for containers

Soft delete enables you to recover containers that were previously marked for deletion. [Learn more](#)

Days to retain deleted containers ⓘ

7

Enable soft delete for file shares

Soft delete enables you to recover file shares that were previously marked for deletion. [Learn more](#)

Days to retain deleted file shares ⓘ

7

**Review**

< Previous

Next : Encryption >

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

Download a template for automation

finalprojadlsstorage\_1665040003505 | Overview

Your deployment is complete

Deployment name: finalprojadlsstorage\_1665040003505  
 Subscription: Azure Pass - Sponsorship  
 Resource group: vp06102022

Start time: 10/6/2022, 12:36:53 PM  
 Correlation ID: 74cbea1b-b064-4098-adbf-28d2fbfa1145

Deployment details

Next steps

Give feedback

Tell us about your experience with deployment

Cost Management

Microsoft Defender for Cloud

● **Create an Azure SQL Database.**

Microsoft services

SQL databases

Create a resource Storage accounts SQL databases Azure DataBricks Quickstart Center Virtual machines More services

Resources

Recent Favorite

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 \*  Resource group \*

**Database details**

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

**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 \***  .database.windows.net

**Location \***

**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 \***

**Password \***

**Confirm password \***   Password and confirm password must match.

**OK**

## Configure

[Feedback](#)

### Service and compute tier

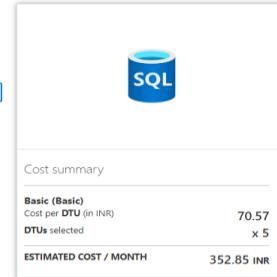
Select from the available tiers based on the needs of your workload. The vCore model provides a wide range of configuration controls and offers Hyperscale and Serverless to automatically scale your database based on your workload needs. Alternately, the DTU model provides set price/performance packages 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)

[Apply](#)

## Create SQL Database

Microsoft

[Basics](#) [Networking](#) [Security](#) [Additional settings](#) [Tags](#) [Review + create](#)

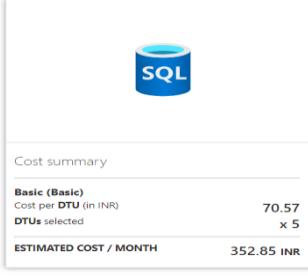
Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'vserverfinalpro' and all databases it manages. [Learn more](#)

### 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 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

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

## 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.

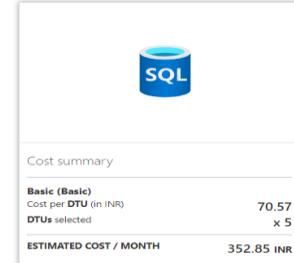
Enable Microsoft Defender for SQL \*  Start free trial  
 Not now

### 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](#)

[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?

Cost summary

| Basic (Basic)          | 70.57 |
|------------------------|-------|
| Cost per DTU (in INR)  | x 5   |
| DTUs selected          |       |
| 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                                                                   |
|----------------------|------------------------|----------------------------------------------------------------------------|
| <input type="text"/> | <input type="text"/> : | <input type="button" value="2 selected"/> <input type="button" value="▼"/> |

Cost summary

| Basic (Basic)          | 70.57 |
|------------------------|-------|
| Cost per DTU (in INR)  | x 5   |
| DTUs selected          |       |
| 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)          | 70.57 |
|------------------------|-------|
| Cost per DTU (in INR)  | x 5   |
| DTUs selected          |       |
| ESTIMATED COST / MONTH |       |
| 352.85 INR             |       |

< Previous Download a template for automation

Home > Microsoft.SQLDatabase.newDatabaseNewServer\_9c1711eda4b1405281013 | Overview

Your deployment is complete

Deployment name: Microsoft.SQLDatabase.newDatabaseNewServer\_9c1711eda4b1405281013  
Subscription: Azure Pass - Sponsorship  
Resource group: vp06102022

Start time: 10/6/2022, 8:09:14 PM Correlation ID: 2d1b097c-3b87-40f2-a3ee-cb17e0037cb3

Deployment details

Next steps

Give feedback

Tell us about your experience with deployment

Deployment succeeded

Deployment Microsoft.SQLDatabase.newDatabaseNewServer\_9c1711eda4b1405281013... 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 > vpfinalproj (vpserverfinalproj/vpfinalproj) | Overview

SQL databases Default Directory (pandeyaibhav9415@gmail.com...) + Create Reservations ...

Filter for any field... Name ↑

- appdb (vpserver/appdb) ...
- vaibhav77897 (vpserver/vaibhav77897) ...
- vpfinalproj (vpserverfinalproj/vpfinalproj) ...

Overview Activity log Tags Diagnose and solve problems Getting started Query editor (preview) Power Platform Power BI Power Apps Power Automate Settings

This database was just created. Do you need any help getting started?

Essentials

Resource group (move) vp06102022 Status Ready Location East US Subscription (move) Azure Pass - Sponsorship Subscription ID 00b9cf4b-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

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

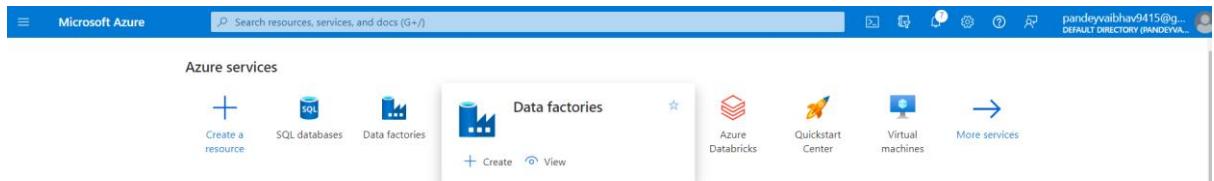
| Rule name    | Start IPv4 address | End IPv4 address |
|--------------|--------------------|------------------|
| firewallrule | 0.0.0.0            | 255.255.255.255  |

Exceptions

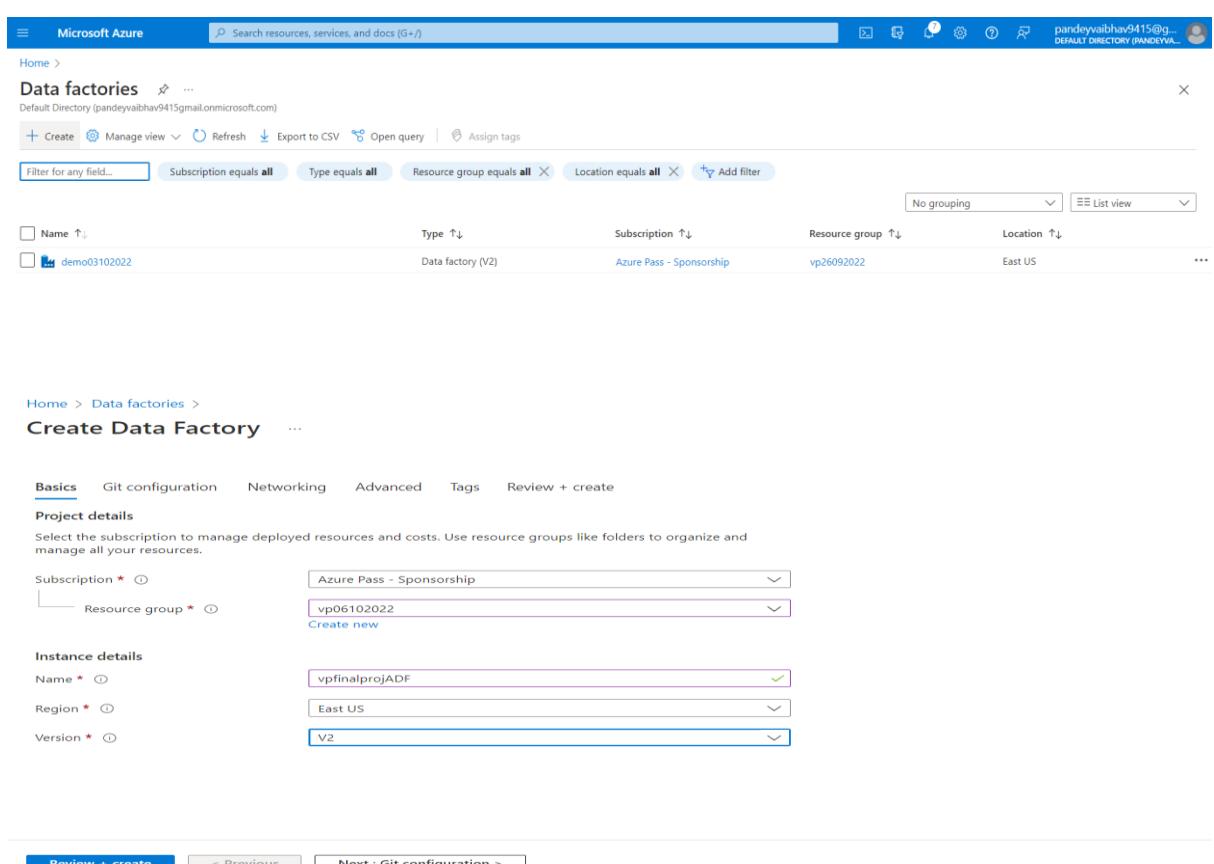
Allow Azure services and resources to access this server

Save Discard

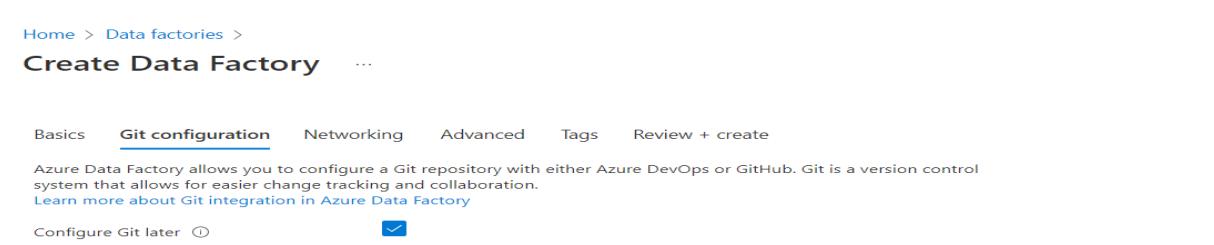
## ● Create a data factory.



The screenshot shows the Microsoft Azure portal interface. In the top navigation bar, the user is in the 'Microsoft Azure' account, with the email 'pandeyvaibhav9415@gmail.com' and the 'DEFAULT DIRECTORY (PANDEYVA...)' selected. The search bar at the top center contains the placeholder 'Search resources, services, and docs (G+/-)'. Below the search bar, there are several icons for Azure services: 'Create a resource' (plus sign), 'SQL databases', 'Data factories' (highlighted with a yellow box), 'Azure Databricks', 'Quickstart Center', 'Virtual machines', and 'More services' (arrow icon). The main content area is titled 'Azure services' and 'Data factories'. It shows a list of recent and favorite resources, with 'Name' as the primary column. A single item is listed: 'demo03102022' (Type: Data factory (V2), Subscription: Azure Pass - Sponsorship, Resource group: vp26092022, Location: East US). The bottom of the page has filter and sorting options, and a 'List view' button.

The screenshot shows the 'Create Data Factory' wizard, Step 1: Basics. The top navigation bar shows 'Home > Data factories > Create Data Factory'. The page title is 'Create Data Factory'. The tabs at the top are 'Basics', 'Git configuration', 'Networking', 'Advanced', 'Tags', and 'Review + create'. The 'Basics' tab is selected. The 'Project details' section asks to select a subscription and resource group. The 'Subscription' dropdown is set to 'Azure Pass - Sponsorship' and the 'Resource group' dropdown is set to 'vp06102022'. The 'Instance details' section includes fields for 'Name' (set to 'vpfinalprojADF'), 'Region' (set to 'East US'), and 'Version' (set to 'V2'). At the bottom, there are buttons for 'Review + create', '< Previous', and 'Next : Git configuration >'.

The screenshot shows the 'Create Data Factory' wizard, Step 2: Git configuration. The top navigation bar shows 'Home > Data factories > Create Data Factory'. The page title is 'Create Data Factory'. The tabs at the top are 'Basics', 'Git configuration', 'Networking', 'Advanced', 'Tags', and 'Review + create'. The 'Git configuration' tab is selected. The 'Git configuration' section explains that Azure Data Factory allows configuration via Git repository (either Azure DevOps or GitHub). It includes a link to 'Learn more about Git integration in Azure Data Factory'. There is a checkbox labeled 'Configure Git later' which is checked. At the bottom, there are buttons for 'Review + create', '< Previous', and 'Next : Review + create >'.

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                               |
|----------------------|---------|----------------------------------------|
| <input type="text"/> | :       | <input type="text"/> 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 >

Home > Microsoft.DataFactory-20221006201613 | Overview X ...

Deployment

Search >Delete Cancel Redeploy Download Refresh

**Overview**

Your deployment is complete

Deployment name: Microsoft.DataFactory-20221006201613  
Subscription: Azure Pass - Sponsorship  
Resource group: vp06102022

Start time: 10/6/2022, 8:41:29 PM  
Correlation ID: 53fe15f5-1f6d-4bee-8949-b86a1809bda Copy

Deployment succeeded Deployment 'Microsoft.DataFactory-20221006201613' to resource group 'vp06102022' was successful.

Pin to dashboard Go to resource group

**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  
[Get started with Microsoft Defender for Cloud >](#)

## ● Configure Databricks cluster

The screenshot shows the Microsoft Azure portal interface. At the top, there's a search bar and a navigation bar with links like 'Gmail', 'YouTube', 'Maps', 'My tabs', 'New folder', 'VP bookmarks', 'Competitive Program...', 'Assignment Submissions', 'Azure Dev - Google...', 'Home - Microsoft A...', and user information 'pandeyvaibhav9415@gmail.com'.

The main area displays the 'Azure services' dashboard with icons for 'Create a resource', 'Data factories', 'SQL databases', 'Storage accounts', 'Azure Synapse Analytics', 'Subscriptions', and 'Databricks'. A callout box highlights the 'Databricks' icon.

Below the dashboard, there's a 'Resources' section with tabs for 'Recent' and 'Favorite'. It shows a table of resources, with one entry for 'vpitestspark' selected. The table includes columns for Name, Type, Resource group, Location, and Subscription.

The URL in the browser is 'portal.azure.com/#home'.

**Create an Azure Databricks workspace**

**Basics**   **Networking**   Advanced   Tags   Review + create

**Project Details**

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

Subscription \*    Resource group \*    Create new

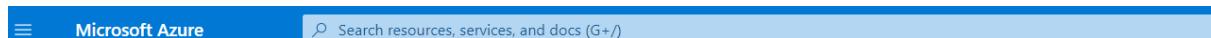
**Instance Details**

Workspace name \*    Region \*    Pricing Tier \*

**Networking**

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)  Yes  No

Deploy Azure Databricks workspace in your own Virtual Network (VNet)  Yes  No



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 | vptfinalprojspark        |
| 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](#)



Home >

vp06102022\_vptfinalprojspark | Overview

pandeyalibhav9415@g... DEFAULT DIRECTORY (PANDEYALIBHAV9415)

Deployment

Search

Delete

Cancel

Redeploy

Download

Refresh



Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name: vp06102022\_vptfinalprojspark  
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.

**Azure services**

- + Create a resource
- Azure Synapse Analytics** (highlighted with a red circle)
- SQL databases
- Subscriptions
- Azure Databricks
- Data factories
- Storage accounts
- Quickstart Center
- Virtual machines
- More services

**Resources**

Recent    Favorite

| Name           | Type              | Last Viewed              |
|----------------|-------------------|--------------------------|
| vpfirstexample | Synapse workspace | vp26092022               |
|                |                   | East US                  |
|                |                   | Azure Pass - Sponsorship |

**Azure Synapse Analytics**

Default Directory (pandeyvaibhav9415@gmail.onmicrosoft.com)

+ Create    Manage view    Refresh    Export to CSV    Open query    Assign tags

Filter for any field...    Subscription equals all    Resource group equals all    Location equals all    Add filter

No grouping    List view

| Name ↑         | Type ↑            | Resource group ↑ | Location ↑ | Subscription ↑           |
|----------------|-------------------|------------------|------------|--------------------------|
| vpfirstexample | Synapse workspace | vp26092022       | East US    | Azure Pass - Sponsorship |

**Create Synapse workspace**

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

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

**Project details**

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

|                        |                                   |
|------------------------|-----------------------------------|
| Subscription *         | Azure Pass - Sponsorship          |
| Resource group *       | vp06102022                        |
| Managed resource group | Enter managed resource group name |

**Workspace details**

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

|                                  |                                                                                           |
|----------------------------------|-------------------------------------------------------------------------------------------|
| Workspace name *                 | vpfinalprojsynapse                                                                        |
| Region *                         | East US                                                                                   |
| Cloud File Lake Creation Plan? * | <input checked="" type="radio"/> From subscription <input type="radio"/> Manually via URL |

**Review + create**    < Previous    Next: Security >

**Create Synapse workspace**

**Region \***

East US

**Select Data Lake Storage Gen2 \***

From subscription     Manually via URL

|                    |                          |
|--------------------|--------------------------|
| Account name *     | (New) vpfinalproj0610    |
| File system name * | (New) vpfinalprojsynapse |

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

**Learn more**

We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the **Storage Blob Data Contributor** role. To enable other users to use this storage account after you create your workspace, perform these tasks:

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

**Review + create**    < Previous    Next: Security >

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. ⓘ

**i** The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a storage account managed via URL under Basics tab. [Learn more](#)

**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

**i** 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. 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 >

Home > Azure Synapse Analytics >  
**Create Synapse workspace** ...

Validation succeeded

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

**Product Details**

Azure Synapse Analytics workspace by Microsoft Serverless SQL est. cost/TB ⓘ **360.23 INR**  
Terms of use | Privacy policy

**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                  |

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

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

Home > Microsoft.Azure.SynapseAnalytics-20221006223145 | Overview

Deployment

Overview

Your deployment is complete

Deployment name: Microsoft.Azure.SynapseAnalytics-20221006223... Start time: 10/6/2022, 10:41:16 PM  
Subscription: Azure Pass - Sponsorship Correlation ID: 90d94529-c2eb-4f8-bf47-c8d3b59265e9

Inputs Outputs Template

Deployment details Next steps Go to resource group

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 >

For making a dedicated SQL Pool

Home > Azure Synapse Analytics

Default Directory (pandeyvaibhav9415@gmail.onmicrosoft.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags

Filter for any field... Subscription equals all Resource group equals all Location equals all Add filter

No grouping List view

| Name               | Type              | Resource group | Location | Subscription             |
|--------------------|-------------------|----------------|----------|--------------------------|
| vpfinalprojsynapse | Synapse workspace | vp06102022     | East US  | Azure Pass - Sponsorship |
| vfirstexample      | Synapse workspace | vp26092022     | East US  | Azure Pass - Sponsorship |

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 JSON View

**Overview**

**Essentials**

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.azuresynthesize.net/workspace=%2fsubscriptions%2f08b9cf4b-aebb-48b5-a74d-f1724247b371>  
Tags (edit) : Click here to add tags

**Analytics pools**

SQL pools  
 Apache Spark pools  
 Data Explorer pools (preview)

**Security**

Encryption

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 \*

vpfinalprosqlpool



Performance level



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\_CI\_AS. [Learn more](#)

Collation \*

SQL\_Latin1\_General\_CI\_AS

[Review + create](#)

[< Previous](#)

[Next: Tags >](#)

Microsoft Azure Search resources, services, and docs (G+) pandeyvaibhav9415@g... DEFAULT DIRECTORY (PANDEYVAI...)

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. I 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\_CI\_AS

[Create](#)

[< Previous](#)

[Download a template for automation](#)

Microsoft Azure Search resources, services, and docs (G+) pandeyvaibhav9415@g... DEFAULT DIRECTORY (PANDEYVAI...)

Home > Microsoft.Azure.Synapse.SqlPoolOnExistingWorkspace\_76ec267b1cf74 | Overview ...

Deployment

Search Delete Cancel Redeploy Download Refresh

Overview

Your deployment is complete

Deployment name: Microsoft.Azure.Synapse.SqlPoolOnExistingWor... Start time: 10/6/2022, 11:15:22 PM  
Subscription: Azure Pass - Sponsorship 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 Flowlet Copy Data tool

Select an item

Use the resource explorer to select or create a new item

Microsoft Azure

Search resources, services, and docs (G+/)

Home > Azure Synapse Analytics > vpfinalprojsynapse

vpfinalprojsynapse

Synapse workspace

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

JSON View

**Essentials**

Resource group (move) : vp06102022

Status : Succeeded

Location : East US

Subscription (move) : Azure Pass - Sponsorship

Subscription ID : 009cf4b-aebb-48b5-a74d-f1724247b371

Managed virtual network : No

Managed identity object id : 7b6a6434-5df9-4f18-b782-7ef01107fe2c

Workspace web URL : <https://web.azuresynthesize.net?workspace=%2bsubscriptions%2c>

Tags (edit) : Click here to add tags

Networking

Primary ADLS Gen2 account : <https://vpfinalproj0610.dfs.core.windows.net>

Primary ADLS Gen2 file system : vpfinalprojsynapse

SQL admin username : sqladminuser

SQL Active Directory admin : [live.com#pandeyvaibhav9415@gmail.com](mailto:pandeyvaibhav9415@gmail.com)

Dedicated SQL endpoint : vpfinalprojsynapse.sql.azuresynapse.net

Serverless SQL endpoint : vpfinalprojsynapse-on-demand.sql.azuresynapse.net

Development endpoint : <https://vpfinalprojsynapse.dev.azuresynthesize.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 \* sparkpoolfinalalp

Isolated compute  Enabled  Disabled

Node size family Memory Optimized

Node size \* Small (4 vCores / 32 GB)

Autoscale \*  Enabled  Disabled

Number of nodes \* 3

Dynamically allocate executors  Enabled  Disabled

Estimated price : 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

... [Edit](#) [Delete](#) [Cancel](#)

\* 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          |
|----------------------|---------|-------------------|
| <input type="text"/> | :       | Apache Spark pool |

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

Home > Azure Synapse Analytics > vpfinalprojsynapse >

## New Apache Spark pool

... [Edit](#) [Delete](#) [Cancel](#)

Validation succeeded

\* Basics \* Additional settings Tags **Review + create**

**Product Details**

Azure Synapse Analytics Apache Spark pool by Microsoft **Est. cost per hour**  
119.31 to 119.31 INR [View pricing details](#)

[Terms of use](#) | [Privacy policy](#)

**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](#) [Delete](#) [Cancel](#) [Redeploy](#) [Download](#) [Refresh](#)

Your deployment is complete

Deployment name: Microsoft.Azure.SynapseAnalytics.SparkPool-20... Start time: 10/7/2022, 9:51:38 AM  
Subscription: Azure Pass - Sponsorship Correlation ID: bccb021-5c20-4d03-8b70-0b23684b31f8 [Copy](#)

Resource group: vpf06102022

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.**

The screenshot shows the Azure Storage Accounts blade. At the top, there are navigation links: Home >, Storage accounts, and a search bar. Below the header are several buttons: + Create, Restore, Manage view, Refresh, Export to CSV, Open query, Assign tags, and Delete. There are also filter buttons for Subscription, Resource group, and Location, along with an 'Add filter' button. The main area displays a table of storage accounts. The columns are: Name, Type, Kind, Resource group, Location, and Subscription. The table lists several accounts, including 'finalprojadlsstorage' which is highlighted with a yellow background. The 'finalprojadlsstorage' row has a 'View' and 'Delete' button at the bottom. The table has a 'No grouping' dropdown and a 'List view' button.

| Name                   | Type            | Kind        | Resource group                      | Location | Subscription             |
|------------------------|-----------------|-------------|-------------------------------------|----------|--------------------------|
| adlsdemo27092022       | Storage account | StorageV2   | vp26092022                          | East US  | Azure Pass - Sponsorship |
| dbstoragesunivab4hyf5q | Storage account | BlobStorage | databricks-rg-vpfinalprojspark-d--  | East US  | Azure Pass - Sponsorship |
| dbstoragejwvz4d57i2562 | Storage account | BlobStorage | databricks-rg-vptestspark-hdiutfz-- | East US  | Azure Pass - Sponsorship |
| finalprojadlsstorage   | Storage account | StorageV2   | vp06102022                          | East US  | Azure Pass - Sponsorship |
| gelstorage12345677     | Storage account | StorageV2   | vp26092022                          | East US  | Azure Pass - Sponsorship |
| vpfinalproj0610        | Storage account | StorageV2   | vp06102022                          | East US  | Azure Pass - Sponsorship |
| vpforsynapse           | Storage account | StorageV2   | vp26092022                          | East US  | Azure Pass - Sponsorship |

Home > Storage accounts > finalprojadlsstorage

**finalprojadlsstorage** Storage account

Search

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser

**Data storage**

- Containers** (highlighted with a red circle)
- File shares
- Queues
- Tables

**Essentials**

|                       |                                        |                    |                                   |
|-----------------------|----------------------------------------|--------------------|-----------------------------------|
| Resource group (move) | : vp06102022                           | Performance        | : Standard                        |
| Location              | : East US                              | Replication        | : Locally-redundant storage (LRS) |
| Subscription (move)   | : Azure Pass - Sponsorship             | Account kind       | : StorageV2 (general purpose v2)  |
| Subscription ID       | : 08b9cf4b-aebb-48b5-a74d-f1724247b371 | Provisioning state | : Succeeded                       |
| Disk state            | : Available                            | 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

Home > Storage accounts > finalprojadlsstorage

**finalprojadlsstorage | Containers**

Search resources, services, and docs (G+)

Search

+ 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   |

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser

**Data storage**

- Containers

**New container**

Name \* adlsfinalsource

Public access level Private (no anonymous access)

Advanced

Create Discard

Home > Storage accounts > finalprojadlsstorage | Containers > adlsfinalsource

**adlsfinalsource** Container

Search

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 |

Home > Storage accounts > finalprojadlsstorage | Containers > adlsfinalsource

**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 |

Upload Completed for movies.csv 4.95 KiB | finalprojadlsstorage

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

Home >

**Data factories** ⚙ ...

Default Directory (pandeyvaibhav9415@gmail.onmicrosoft.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags

Filter for any field... Subscription equals all Type equals all Resource group equals all Location equals all Add filter

No group

| Name ↑↓               | Type ↑↓           | Subscription ↑↓          | Resource group ↑↓ |
|-----------------------|-------------------|--------------------------|-------------------|
| demo03102022          | Data factory (V2) | Azure Pass - Sponsorship | vp26092022        |
| <b>vpfinalprojADF</b> | Data factory (V2) | Azure Pass - Sponsorship | vp06102022        |

View

---

Home > Data factories >

**vpfinalprojADF** Data factory (V2)

Search Delete

Overview Activity log Access control (IAM) Tags Diagnose and solve problems

Settings Networking Managed identities Properties Locks

Getting started Quick start Monitoring

Essentials

|                                                        |                                               |
|--------------------------------------------------------|-----------------------------------------------|
| Resource group (move) : vp06102022                     | Type : Data factory (V2)                      |
| Status : Succeeded                                     | Getting started : <a href="#">Quick start</a> |
| Location : East US                                     |                                               |
| Subscription (move) : Azure Pass - Sponsorship         |                                               |
| Subscription ID : 08b9cf4b-aebb-48b5-a74d-f1724247b371 |                                               |

Getting started

Open Azure Data Factory Studio Start authoring and monitoring your data pipelines and data flows. [Open](#)

Read documentation Learn how to be productive quickly. Explore concepts, tutorials, and samples. [Learn more](#)

Monitoring

---

Microsoft Azure | Data Factory > vpfinalprojADF

Validate all Publish all

Factory Resources Pipelines Datasets Data flows Power Query

Filter resources by name

dataflow1 Validate Data flow debug

Add Source Add Flowlet

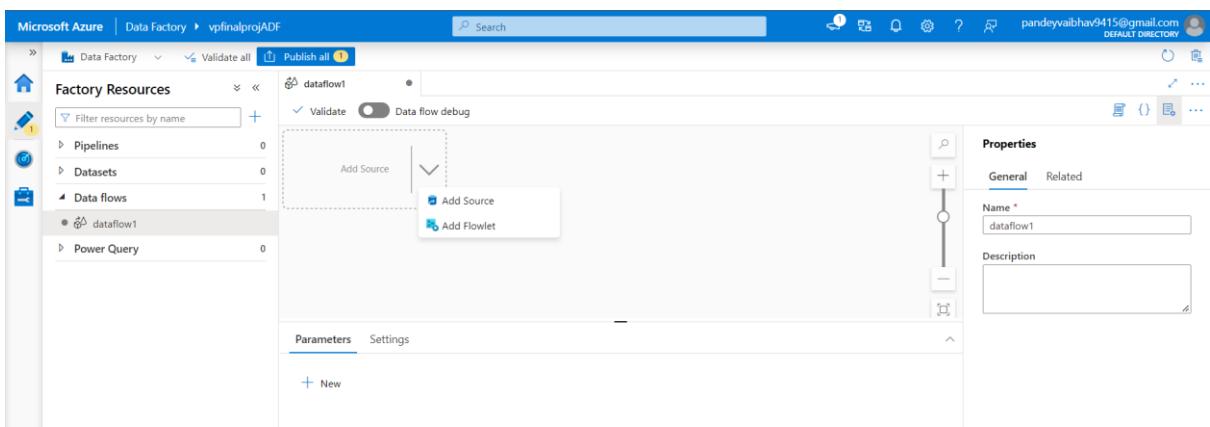
Properties General Related

Name \* dataflow1

Description

Parameters Settings

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 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

**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 finalprojdsstorage

Test connection To linked service To file path

Annotations New

Parameters Advanced

Create Cancel Test connection

**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

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (0), 'Datasets' (1), 'Data flows' (1), and 'Power Query' (0). The main area displays a data flow named 'dataflow1' with one step: 'CleaningGenreRomance'. The properties pane on the right shows the 'General' tab with 'Name' set to 'dataflow1' and 'Description' left empty.

Now go for data cleansing:

The screenshot shows the Microsoft Azure Data Factory Data Flow blade. The 'Source type' dropdown is open, showing options like 'Dataset' (selected) and 'Inline'. The 'Schema modifier' dropdown is also open, with 'Derived Column' highlighted by a red arrow. The main area shows the data flow 'task1dataflow1' with a single step 'CleaningGenreRomance'. A warning message at the bottom says: 'Please turn on debug mode and wait until the cluster is ready to preview data...'.

A modal dialog titled 'Turn on data flow debug' is open on the right. It includes fields for 'Integration runtime' (set to 'AutoResolveIntegrationRuntime'), 'Region' (set to 'AutoResolve'), 'Compute size' (set to 'Small'), and 'Debug time to live' (set to '1 hour'). At the bottom are 'OK' and 'Cancel' buttons.

**Dataflow expression builder**

Derived Columns

+ Create new

abc Genre

Column name \* **Genre**

Expression

```
case(Genre == "romance", "Romance", Genre == "Romance", "Romance", Genre == "Comdy", "Comedy", Genre == "comedy", "Comedy", Genre)
```

Expression elements Expression values

Data preview Refresh

| Output: Genre abc | Genreabc |
|-------------------|----------|
| Romance           | Romance  |
| Comedy            | Comedy   |
| Comedy            | Comedy   |
| Comedy            | Comedy   |

Save and finish Cancel Clear contents

**Data Factory** Validate all Publish all

Factory Resources

- Pipelines
- Datasets
- Data flows
- task1dataflow1
- Power Query

task1dataflow1 for task1source

Validate Data flow debug Debug Settings

Derived column's settings

Output stream name \* **derivedColumn1**

Description **Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide Gross, Year**

Incoming stream \* **CleaningGenreRomance**

Add Source

**Data Factory** Validate all Publish all

Factory Resources

- Pipelines
- Datasets
- Data flows
- task1dataflow1
- Power Query

task1dataflow1 for task1source

Validate Data flow debug Debug Settings

Derived column's settings

Optimize

Partition option \*

Use current partitioning  Single partition  Set partitioning

**Data Factory** Validate all Publish all

Factory Resources

- Pipelines
- Datasets
- Data flows
- task1dataflow1
- Power Query

task1dataflow1 for task1source

Validate Data flow debug Debug Settings

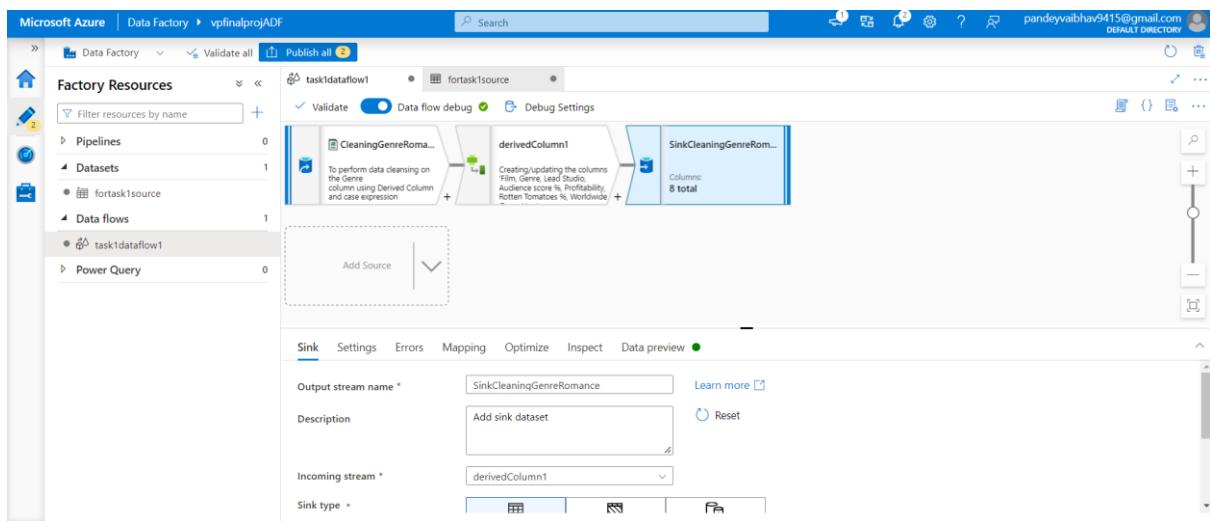
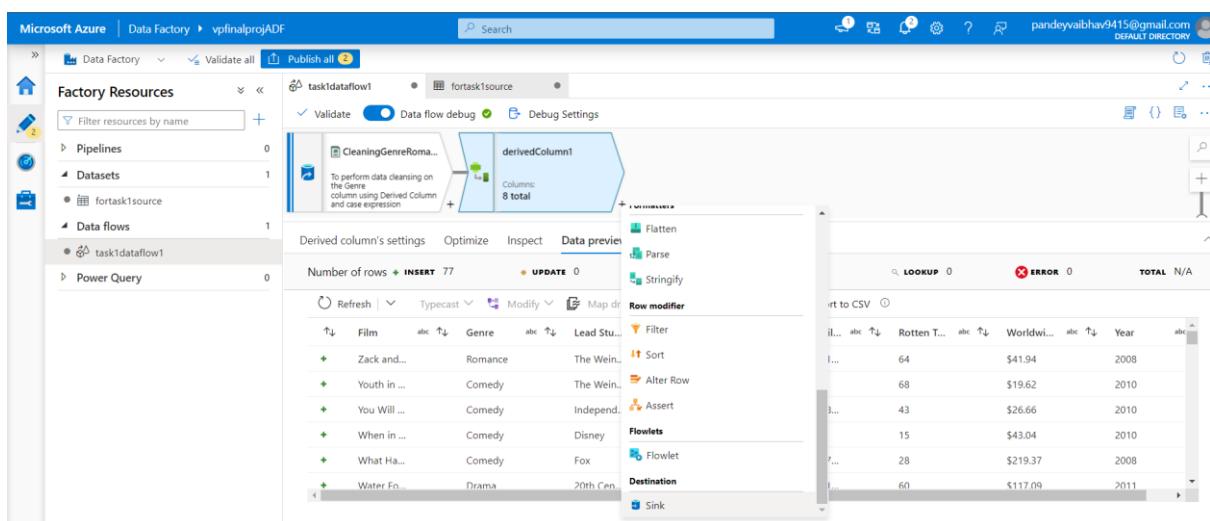
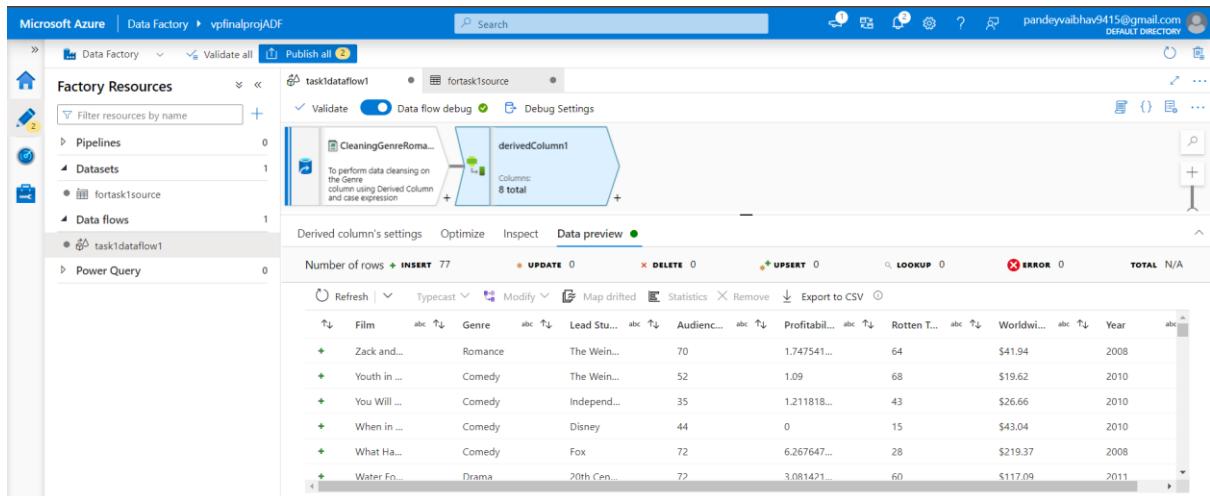
Derived column's settings

Inspect

Schema

Input Output

| Number of columns New* 0 | Updated* 1 | Unchanged 7 | Total 8    |
|--------------------------|------------|-------------|------------|
| Order ↑↓                 | Column ↑↓  | Type ↑↓     | Updated ↑↓ |
| 1 Film                   |            | abc string  |            |
| 2 Genre                  |            | abc string  | Genre      |
| 3 Lead Studio            |            | abc string  |            |
| 4 Audience score %       |            | abc string  |            |
| 5 Profitability          |            | abc string  |            |
| 6 Rotten Tomatoes %      |            | abc string  |            |



Microsoft Azure | Data Factory > vpfinalprojADF

Factory Resources

- Pipelines
- Datasets
- Data flows
- Power Query

task1dataflow1

task1source

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 |

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Sink type \* Dataset Inline

Dataset \* Select... + New

Options Allow schema drift Validate schema

Continue Cancel

Microsoft Azure | Data Factory > vpfinalprojADF

Factory Resources

- Pipelines
- Datasets
- Data flows
- Power Query

task1dataflow1

task1source

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

Object Explorer

Connect > vpfinalproj.database.windows.net (SQL Server)

SQLQuery1.sql - vp...\_pdb (sqladmin (68))

```
ASPIRE varchar(300),
DOORS varchar(300),
BODY varchar(300),
DRIVE varchar(300),
CYLINDERS varchar(300),
HP int,
RPM int,
NPYCITY int,
MPGHY int,
PRICE int)

select * FROM autooutput;
truncate table autooutput;
```

100 % Results Messages

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 >

Object Explorer

Connect > vpfinalproj.database.windows.net

SQLQuery2.sql - vp...\_sqladminfinal (53)\*

```
create table movieoutput
(film varchar(300),
Genre varchar(300),
Lead_Studio varchar(300),
Audience_score_pcnt int,
Profitability float,
Rotten_Tomatoes_pcnt int,
Worldwide_Gross varchar(100),
Year int
);

select * FROM movieoutput;
```

100 % Results Messages

Commands completed successfully.

Completion time: 2022-10-07T12:00:37.8940709+05:30

```
100 % 
Results Messages
Film | Genre | Lead_Studio | Audience_score_prcnt | Profitability | Rotten_Tomatoes_prcnt | Worldwide_Gross | Year
select * FROM movieoutput;
```

Microsoft Azure | Data Factory > vpfinalprojADF

Factory Resources

- Pipelines
- Datasets
- Data flows
- Power Query

task1dataflow1

Validate Data flow debug Debug Settings

CleaningGenreRomance 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 Gross

SinkCleaningGenreRomance Columns: 8 total

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Sink type Dataset

Dataset Select... + New

Options Allow schema drift Validate schema

OK Back Cancel

Set properties

Name: AzureSqlTable1

Linked service: AzureSqlDatabase1

Select from existing table Create new table

Table name: dbo.movieoutput

Import schema: From connection/store None

Microsoft Azure | Data Factory > vpfinalprojADF

Factory Resources

- Pipelines
- Datasets
- AzureSqlTable1
- fortask1source
- Data flows
- task1dataflow1
- Power Query

task1dataflow1

Validate Data flow debug Debug Settings

CleaningGenreRomance 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 Gross

SinkCleaningGenreRomance Columns: 8 total

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

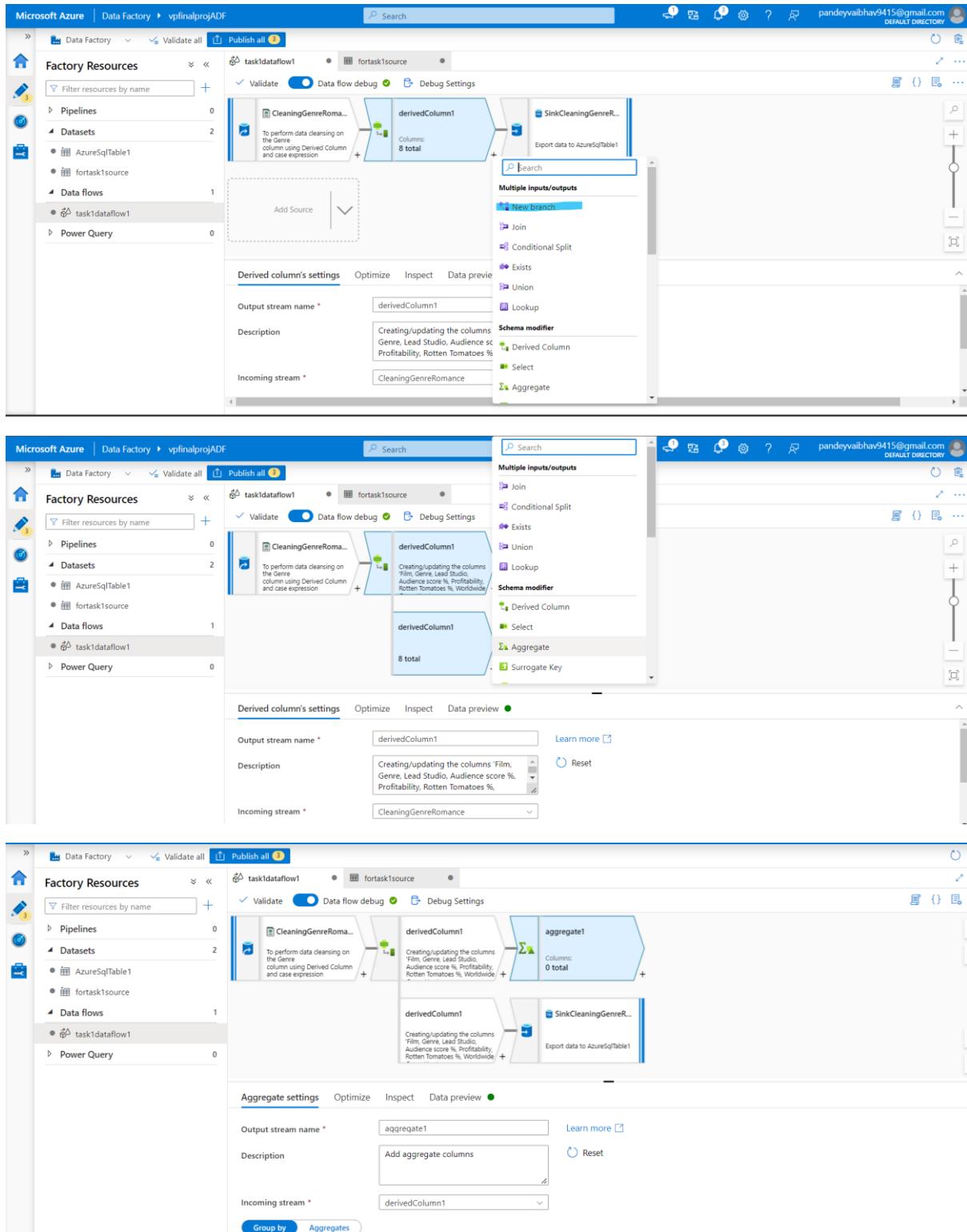
Output stream name: SinkCleaningGenreRomance

Description: Export data to AzureSqlTable1

Incoming stream: derivedColumn1

Sink type Dataset

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

Validate Data flow debug Debug Settings

CleaningGenreRoma... derivedColumn1 CountMoviesBasedOnGenre...

To perform data cleansing on the Genre column using Derived Column and case expression

Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide

derivedColumn1 SinkCleaningGenreR...

Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide

CountMoviesBasedOnGenre...

Columns: 2 total

SinkCleaningGenreR... Export data to AzureSqlTable1

Aggregate settings Optimize Inspect Data preview

Incoming stream \* derivedColumn1

Group by Aggregates

Columns Name as

+abc\_Genre

Genre

Data flows

- task1dataflow1
- Power Query

Add Source

Aggregate settings Optimize Inspect Data preview

Grouped by: Genre

Add Clone Delete Open expression builder

Column Expression

Film count(Film)

Dataflow expression builder

CountMoviesBasedOnGenre

Aggregate Columns

Create new

123 Film

Column name \* Film

Expression count(Film)

Save

Expression elements

- All
- Functions
- Input schema
- Parameters
- Cached lookup
- Data flow library functions

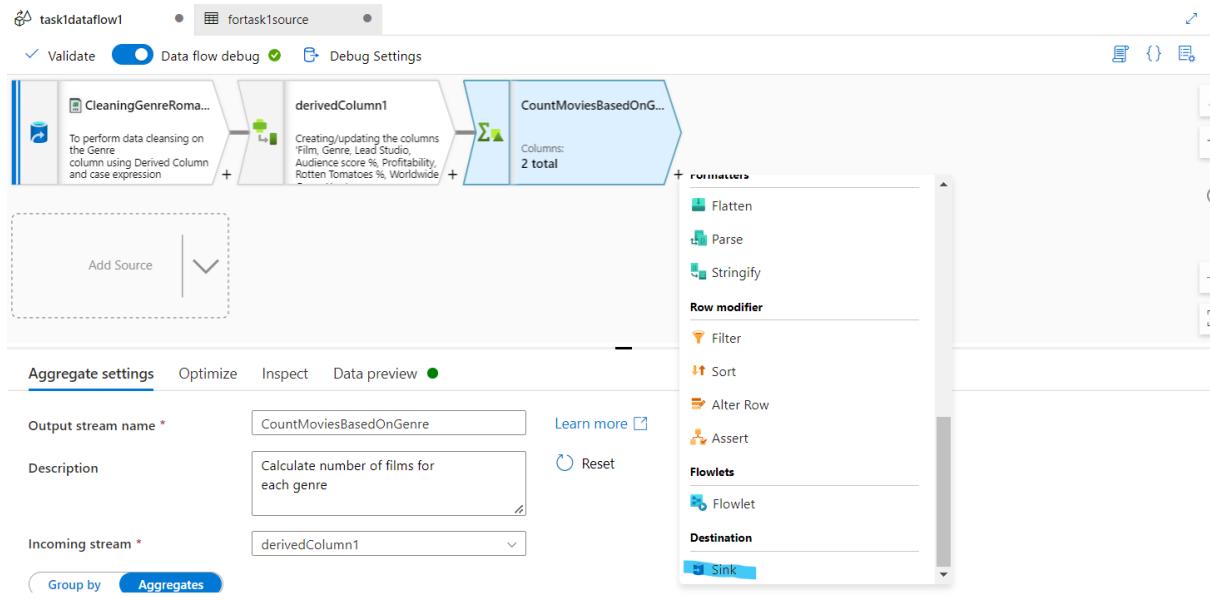
Expression values

Filter by keyword

Create new

abc Film abc Genre abc Lead Studio abc Audience score %

Save and finish Cancel Clear contents



Home > Storage accounts > finalprojadlsstorage | Containers

finalprojadlsstorage | Containers

Search

+ Container Change access level Restore containers Refresh Delete

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)

Create Discard

Home > Storage accounts > finalprojadlsstorage | Containers > solution

solution

Search

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)

Name   Modified   Access tier   Archive status   Blob type   Size   Lease state

|            |  |  |  |  |  |  |
|------------|--|--|--|--|--|--|
| genreCount |  |  |  |  |  |  |
|------------|--|--|--|--|--|--|

Successfully added directory  
Successfully added directory 'genreCount'

**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

Continue Cancel

**Set properties**

Name: DelimitedText1  
Linked service: AzureDataLakeStorage1  
File path: solution / genreCount / File name  
First row as header:   
Import schema: From connection/store  
OK Back Cancel

**Factory Resources**

- Pipelines
- Datasets
  - AzureSqlTable1
  - DelimitedText1
  - fortask1source
- Data flows
  - task1dataflow1
- Power Query

**task1dataflow1**

Validate Data flow debug Debug Settings

CleaningGenreRoma... derivedColumn1 CountMovies...  
To perform data cleansing on the Genre column using Derived Column and case expression  
Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Sink type: Dataset  
Dataset: Select...  
Options: Allow schema drift  Validate schema

**Data preview**

| Genre     | Film |
|-----------|------|
| Romance   | 15   |
| Comedy    | 43   |
| Drama     | 13   |
| Animation | 4    |
| Fantasy   | 1    |
| Action    | 1    |

**Notifications**

Dismiss all

**Publishing completed**  
Successfully published a few seconds ago

**Successfully applied settings**  
Successfully applied settings to AzureDataLakeStorage1 (Linked service). 3 minutes ago

**Factory Resources**

- Pipelines
- Datasets
  - AzureSqlTable1
  - DelimitedText1
  - fortask1source
- Data flows
  - task1dataflow1
- Power Query

**task1dataflow1**

Validate Data flow debug Debug Settings

CleaningGenreRoma... derivedColumn1 CountMovies...  
To perform data cleansing on the Genre column using Derived Column and case expression  
Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Sink type: Dataset  
Dataset: Select...  
Options: Allow schema drift  Validate schema

**Notifications**

Dismiss all

**Publishing completed**  
Successfully published a few seconds ago

**Successfully applied settings**  
Successfully applied settings to AzureDataLakeStorage1 (Linked service). 3 minutes ago

**Factory Resources**

- Pipelines
- Datasets
  - AzureSqlTable1
  - DelimitedText1
  - fortask1source
- Data flows
  - task1dataflow1
- Power Query

**task1dataflow1**

Validate Data flow debug Debug Settings

CleaningGenreRoma... derivedColumn1 CountMovies...  
To perform data cleansing on the Genre column using Derived Column and case expression  
Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide

Add Source

Sink Settings Errors Mapping Optimize Inspect Data preview

Sink type: Dataset  
Dataset: Select...  
Options: Allow schema drift  Validate schema

task1q2dataflow1 pipeline1

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

Copy data1

Data flow1

Source

General

Source dataset \*

fortask1source

File path type

File path in dataset

Filter by last modified

Start time (UTC)

End time (UTC)

Recursively

Enable partition discovery

task1dataflow1 pipeline1 pipeline2 dataflow1 task1dataflow1\_copy1

Activities

Move & transform

- Copy data
- Data flow

Batch Service

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Copy data1

Data flow1

Source

General

Source dataset \*

DelimitedText3

File path type

File path in dataset

Filter by last modified

Start time (UTC)

End time (UTC)

Recursively

Enable partition discovery

task1dataflow1 pipeline1 pipeline2 dataflow1 task1dataflow1

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

Copy data1

Data flow1

Sink

General

Sink dataset \*

Select...

Set properties

Name

DelimitedText4

Linked service \*

AzureDataLakeStorage2

File path

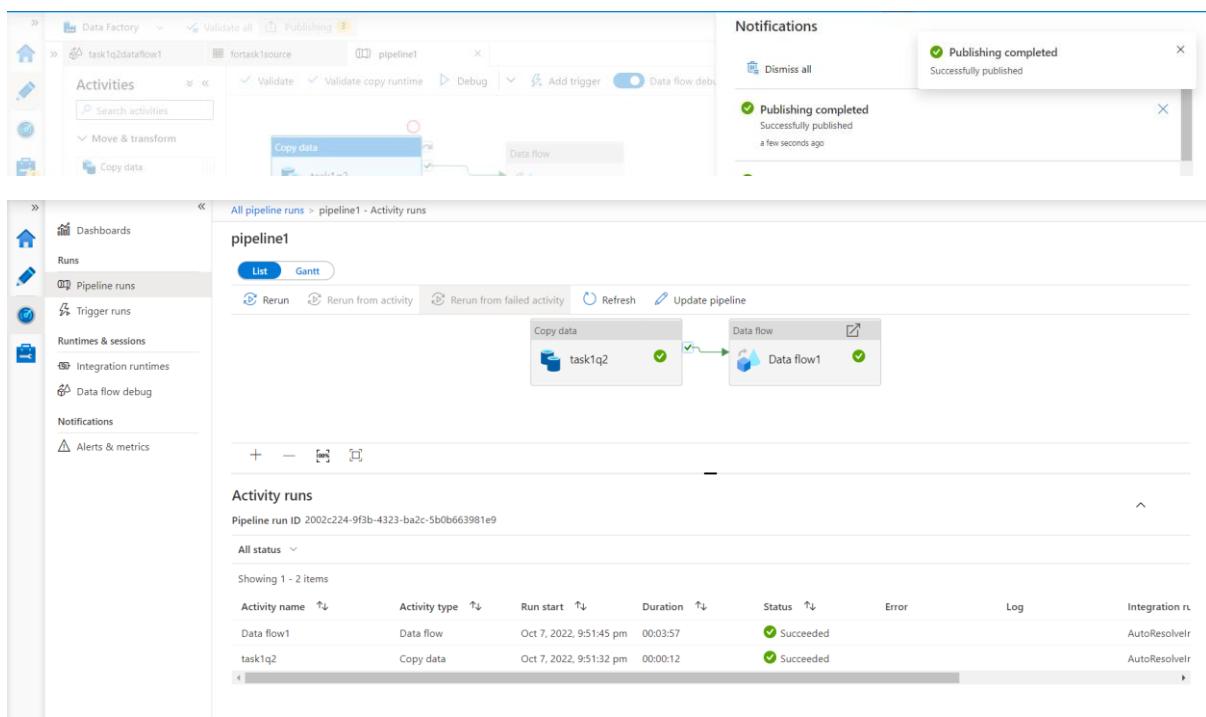
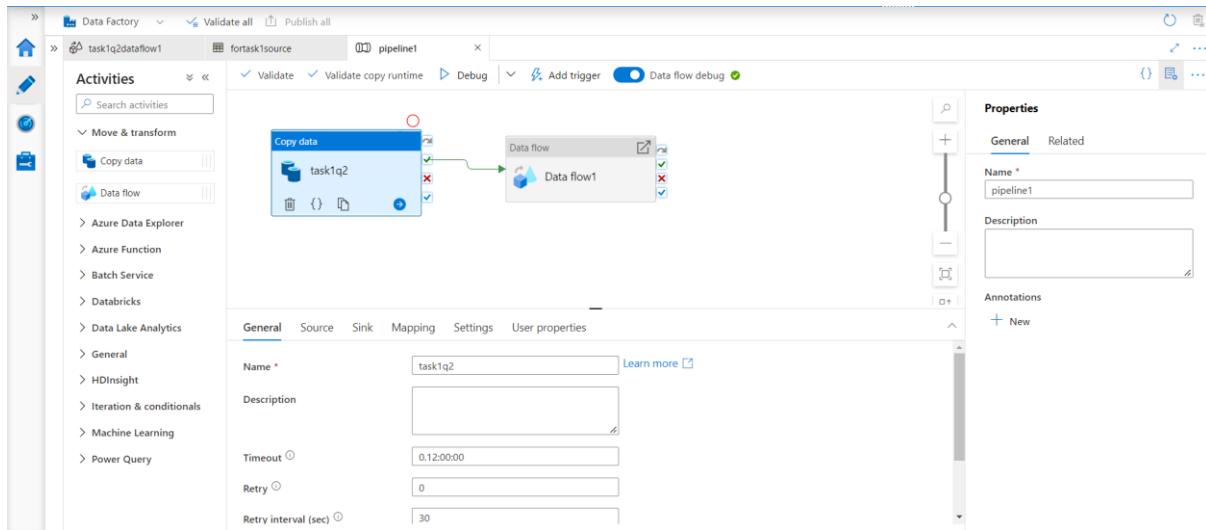
solution / genreCount / File name

First row as header

Import schema

From connection/store

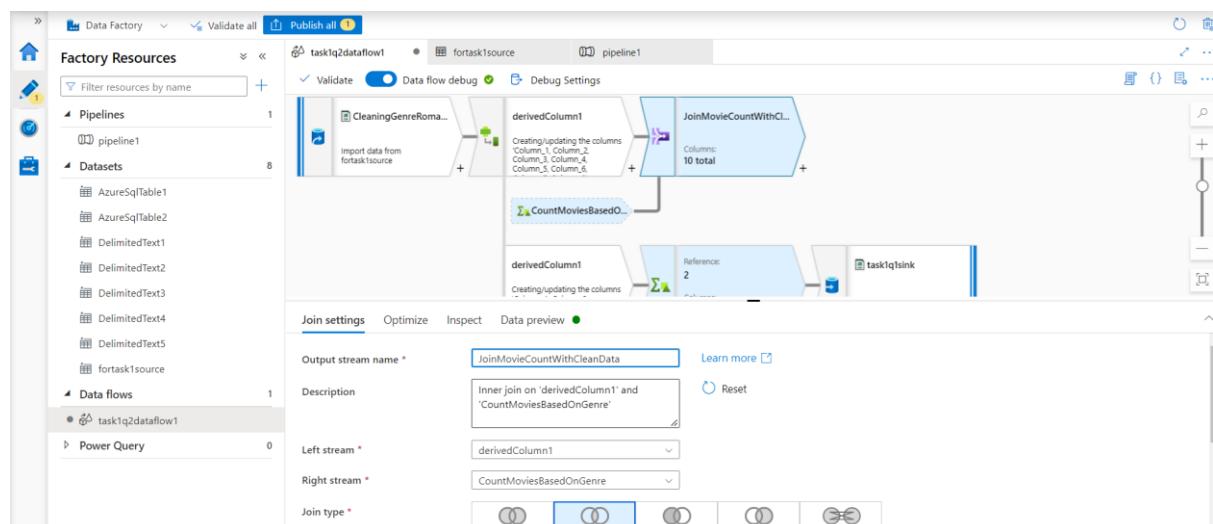
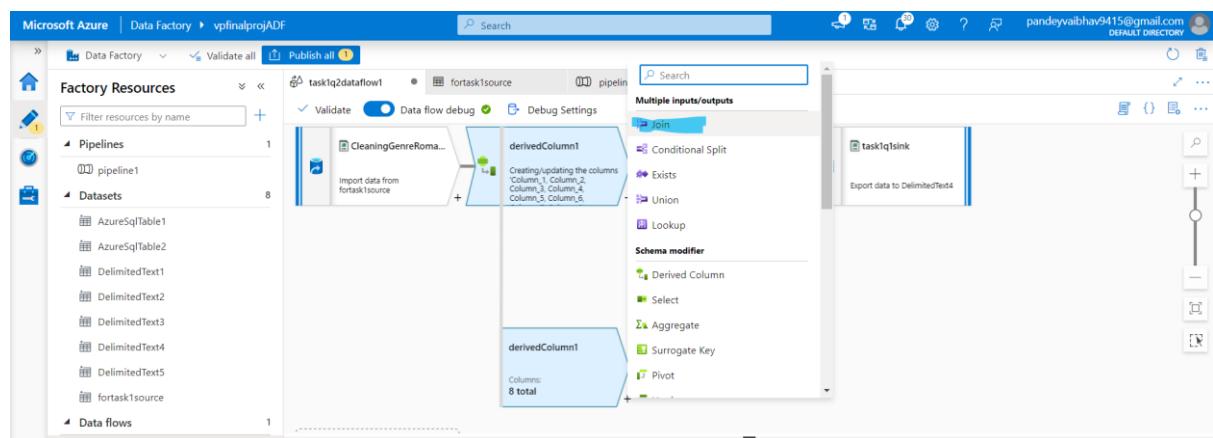
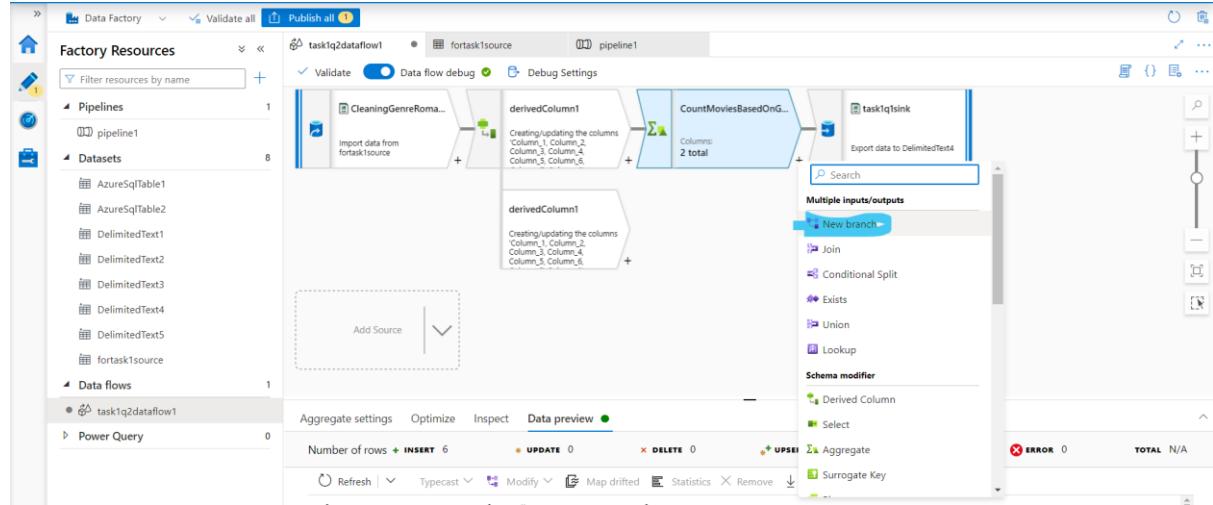
OK Back Cancel

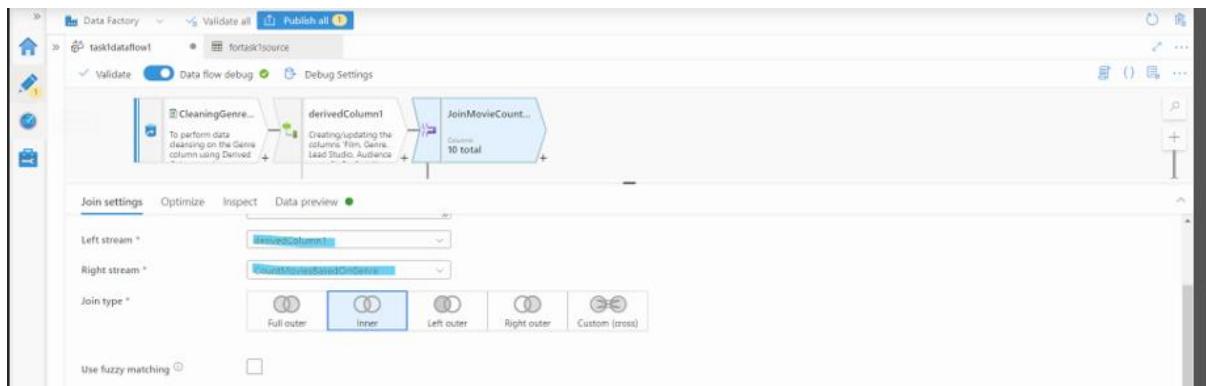


The screenshot shows the Azure Storage Explorer interface. It displays the contents of the 'solution' container under the 'finalprojadlsstorage | Containers' path. The table below shows the list of blobs:

| Name                                            | Modified                | Access tier    | Archive status | Blob type  | Size     | Lease state |
|-------------------------------------------------|-------------------------|----------------|----------------|------------|----------|-------------|
| _SUCCESS                                        | 10/7/2022, 10:10:42 ... | Hot (Inferred) |                | Block blob | 0 B      | Available   |
| movies.csv                                      | 10/7/2022, 10:06:43 ... | Hot (Inferred) |                | Block blob | 6.23 Kib | Available   |
| part-00000-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:38 ... | Hot (Inferred) |                | Block blob | 11 B     | Available   |
| part-00010-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:37 ... | Hot (Inferred) |                | Block blob | 22 B     | Available   |
| part-00045-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:37 ... | Hot (Inferred) |                | Block blob | 20 B     | Available   |
| part-00076-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:37 ... | Hot (Inferred) |                | Block blob | 21 B     | Available   |
| part-00118-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:37 ... | Hot (Inferred) |                | Block blob | 23 B     | Available   |
| part-00178-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:38 ... | Hot (Inferred) |                | Block blob | 21 B     | Available   |
| part-00183-457b4a4a-3cd7-42ee-9f26-4cf508ddb... | 10/7/2022, 10:10:38 ... | Hot (Inferred) |                | Block blob | 20 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.





**Join settings**   **Optimize**   **Inspect**   **Data preview**

**Join type \***    Inner    Full outer    Left outer    Right outer    Custom (cross)

**Use fuzzy matching**

**Combine text parts**

**Similarity score column**

**Join conditions \***   **Left:** derivedColumn1's column   **Similarity Threshold** 80   **Right:** CountMoviesBasedOnGenre's column

abc Genre    abc Genre

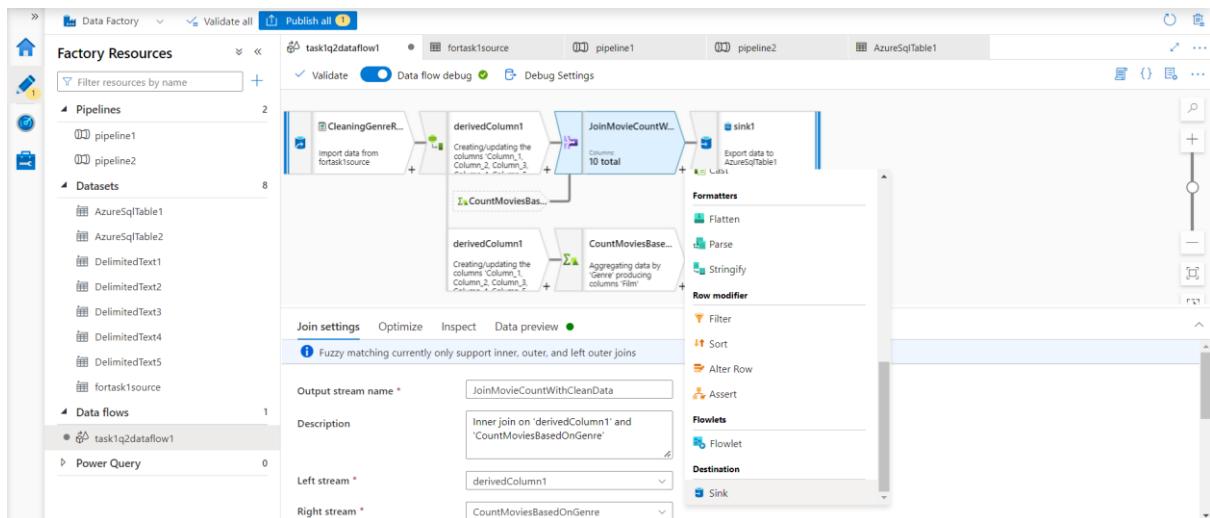
| Number of columns: Left: derivedColumn1 8 |                   | Right: CountMoviesBasedOnGenre 2 |                         | Total 10        |
|-------------------------------------------|-------------------|----------------------------------|-------------------------|-----------------|
| Order ↑↓                                  | Column ↑↓         | Type ↑↓                          | Fed by ↑↓               | Used in Join ↑↓ |
| 1                                         | Film              | abc string                       | CleaningGenreRomance    |                 |
| 2                                         | Genre             | abc string                       | derivedColumn1          | ✓               |
| 3                                         | Lead Studio       | abc string                       | CleaningGenreRomance    |                 |
| 4                                         | Audience score %  | abc string                       | CleaningGenreRomance    |                 |
| 5                                         | Profitability     | abc string                       | CleaningGenreRomance    |                 |
| 6                                         | Rotten Tomatoes % | abc string                       | CleaningGenreRomance    |                 |
| 7                                         | Worldwide Gross   | abc string                       | CleaningGenreRomance    |                 |
| 8                                         | Year              | abc string                       | CleaningGenreRomance    |                 |
| 9                                         | Genre             | abc string                       | CountMoviesBasedOnGenre | ✓               |
| 10                                        | Film              | 12l 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          | abc ↑↓ | Film | 12l ↑↓ | Genre   | abc ↑↓ | Genre   | abc ↑↓ | Lead Studio... | abc ↑↓ | Audien... | abc ↑↓ | Profitabi... | abc ↑↓ | Rotten T... | abc ↑↓ |
|----|---------------|--------|------|--------|---------|--------|---------|--------|----------------|--------|-----------|--------|--------------|--------|-------------|--------|
| +  | Zack and...   | 15     |      |        | Romance |        | Romance |        | The Wei...     |        | 70        |        | 1.747541...  |        | 64          |        |
| +  | Waitress      | 15     |      |        | Romance |        | Romance |        | Independ...    |        | 67        |        | 11.08974...  |        | 89          |        |
| +  | Waiting ...   | 15     |      |        | Romance |        | Romance |        | Independ...    |        | 53        |        | 0.005        |        | 6           |        |
| +  | Tyler Per...  | 15     |      |        | Romance |        | Romance |        | Independ...    |        | 47        |        | 3.7241924    |        | 46          |        |
| +  | Twilight: ... | 15     |      |        | Romance |        | Romance |        | Independ...    |        | 68        |        | 6.383363...  |        | 26          |        |
| +  | Twilight      | 15     |      |        | Romance |        | Romance |        | Summit         |        | 82        |        | 10.18002...  |        | 49          |        |



Sink | Settings | Errors | Mapping | Optimize | Inspect | Data preview

**Sink**

Output stream name \* sink1 | Learn more

Description Export data to AzureSqlTable1 |

Incoming stream \* JoinMovieCountWithCleanData

Sink type \*

|         |        |       |
|---------|--------|-------|
| Dataset | Inline | Cache |
|---------|--------|-------|

Dataset \* AzureSqlTable1 | Test connection | Open | New

Options  Allow schema drift |  Validate schema

Sink | Settings | Errors | **Mapping** | Optimize | Inspect | Data preview

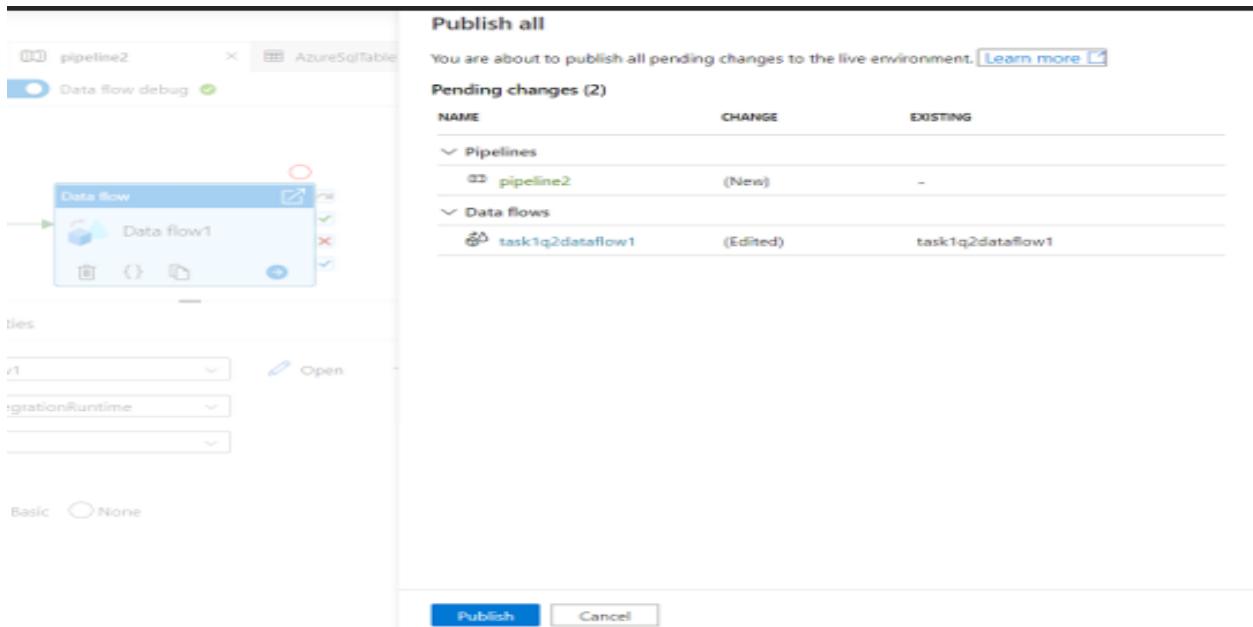
**Mapping**

| Input columns                    | Output columns            |
|----------------------------------|---------------------------|
| CleaningGenreRomance@Film        | abc_Film                  |
| derivedColumn1@Genre             | abc_Genre                 |
| Lead Studio                      | abc_Lead_Studio           |
| Audience score %                 | 123_Audience_score_prcnt  |
| Profitability                    | 123_Profitability         |
| Rotten Tomatoes %                | 123_Rotten_Tomatoes_prcnt |
| Worldwide Gross                  | abc_Worldwide_Gross       |
| Year                             | 123_Year                  |
| 123_CountMoviesBasedOnGenre@Film | 123_FilmCount             |

| Sink                                     | Settings          | Errors            | Mapping           | Optimize          | Inspect           | Data preview     | ●                    |               |        |                      |                      |      |         |           |         |
|------------------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|------------------|----------------------|---------------|--------|----------------------|----------------------|------|---------|-----------|---------|
| Number of rows                           | + <b>INSERT</b> 0 | ● <b>UPDATE</b> 0 | ✗ <b>DELETE</b> 0 | + <b>UPSERT</b> 0 | 🔍 <b>LOOKUP</b> 0 | ✗ <b>ERROR</b> 0 | TOTAL 77             |               |        |                      |                      |      |         |           |         |
| ↻ Refresh   ⏺ Statistics ⏹ Export to CSV |                   |                   |                   |                   |                   |                  |                      |               |        |                      |                      |      |         |           |         |
| ↑ ↓                                      | Film              | abc ↑ ↓           | Genre             | abc ↑ ↓           | Lead_Studio       | abc ↑ ↓          | Audience_... 123 ↑ ↓ | Profitability | 12 ↑ ↓ | Rotten_To... 123 ↑ ↓ | Worldwide... abc ↑ ↓ | Year | 123 ↑ ↓ | FilmCount | 123 ↑ ↓ |
| +                                        | 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      |           |         |

The screenshot shows the Azure Data Factory Data Flow blade. It displays two parallel data flow pipelines. The top pipeline consists of a 'CleaningGenre...' component, a 'derivedColumn1' component (Creating/Updating the columns Film, Genre, Lead Studio, Audience), and a 'JoinMovieCount...' component (Joining 10 total). The bottom pipeline consists of a 'derivedColumn1' component, a 'CountMoviesBas...' component (Calculate number of films for each genre), and a 'Task1sink1' component (Export data to DelimitedText1). Each pipeline has its own 'Data preview' section at the bottom, showing sample data and row counts.

The screenshot shows the Azure Data Factory interface. At the top, there are navigation links for 'Validate' and 'Debug', followed by a 'Data flow debug' toggle switch which is currently off. Below the header is a 'Data flow' panel containing a single item named 'Data flow1'. The main workspace below the panel is currently empty. On the right side, there is a vertical toolbar with icons for creating new items like 'Data flow', 'Pipeline', 'Dataset', and 'Mapping'. The bottom of the screen features a navigation bar with tabs: 'General', 'Settings' (which is selected), 'Parameters', and 'User properties'. Under the 'Settings' tab, there are several configuration options: 'Data flow' set to 'task1q2dataflow1', 'Run on (Azure IR)' set to 'AutoResolveIntegrationRuntime', 'Compute size' set to 'Small', and a 'Logging level' section where 'Verbose' is selected. There are also sections for 'Advanced' and 'Sink properties'.



**Data Factory** Validate all Publish all

**Activities**  Add trigger  Data flow debug

Trigger now Trigger on-demand run of the last published pipeline

**All pipeline runs > pipeline2 - Activity runs**

**Activity runs**

Pipeline run ID 3712a2a1-5a58-460c-b0b1-a6bf7674c084

| Activity name | Activity type | Run start                | Duration | Status    | Error | Log | Integration r... |
|---------------|---------------|--------------------------|----------|-----------|-------|-----|------------------|
| Data flow1    | Data flow     | Oct 7, 2022, 11:49:53 pm | 00:09:13 | Succeeded |       |     | AutoResolve...   |
| task1q2       | Copy data     | Oct 7, 2022, 11:49:42 pm | 00:00:10 | Succeeded |       |     | AutoResolve...   |

Home > Storage accounts > finalprojadlsstorage | Containers >

**solution** Container

Search  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 | Block type | Size | Lease state |
|------------|-----------------------|----------------|----------------|------------|------|-------------|
| GenreCount | 10/8/2022, 3:14:40 PM | Hot (Inferred) |                | Block blob | 72 B | Available   |

**Settings**

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

The screenshot shows the SQL Server Management Studio interface. The Object Explorer on the left lists the database structure for 'vpserverfinalproj.database.windows.net'. A query window in the center contains T-SQL code to create a table 'movieoutput' and insert data from the 'titles' table, followed by a truncate command. The results grid below shows 16 movie records.

```
Object Explorer
Connect C \v
SQLQuery2.sql - vp$sqldminifinal (53)* - SQLQuery1.sql - vp...pdb (sqladmin (66)) - -vs550.sql - vpfir... (sqladminuser (74)) - -vsCC70.sql - vpfir... (sqladminuser (72))
[SQLQuery2.sql - vp$sqldminifinal (53)*] 4 SQLQuery1.sql - vp...pdb (sqladmin (66)) * -vs550.sql - vpfir... (sqladminuser (74)) - -vsCC70.sql - vpfir... (sqladminuser (72))

-- Create movieoutput table
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,
    FileCount int
);

-- Insert data from titles
SELECT * FROM movieoutput;
TRUNCATE TABLE movieoutput;

Results 1 Messages
Film Genre Lead_Studio Audience_score_prcnt Profitability Rotten_Tomatoes_prcnt Worldwide_Gross Year FileCount
1 Zack and Miri Make a Porno Romance The Weinstein Company 70 1.747541667 64 $41.94 2008 15
2 Wall-E Romance Independent 67 11.0897415 89 $22.13 2008 15
3 The Polar Express Romance Independent 53 1.0000 6 $0.01 2004 15
4 Tyler Perry's Why Did I Get Married Romance Independent 47 3.7241924 46 $55.96 2007 15
5 Twilight Breaking Dawn Romance Independent 68 6.383363084 26 $70.17 2011 15
6 Twilight Romance Summit 82 10.180027486 49 $376.66 2008 15
7 Something Borrowed Romance Independent 48 1.719512468 15 $60.18 2011 15
8 P.S. I Love You Romance Independent 82 5.103116833 21 $153.09 2007 15
9 One Day Romance Independent 54 3.682733333 37 $55.24 2011 15
10 New Year's Eve Romance Warner Bros. 48 2.536426571 8 $142.04 2011 15
11 Music and Lyrics Romance Warner Bros. 70 3.6741055 63 $145.90 2007 15
12 Monte Carlo Romance 20th Century Fox 50 1.9832 38 $39.66 2011 15
13 Midnight in Paris Romance Sony 84 8.744705882 93 $148.66 2011 15
14 The Curious Case of Jane Eyre Romance Universal 77 0 85 $37.13 2011 15
15 The Amazing Universe Romance Independent 64 0.652603178 44 $39.37 2007 15
16 The Curious Case of Benjamin Button Fantasy Warner Bros. 81 1.78394375 73 $205.43 2008 1
17 Inter-En-Elephantine Persian 20th Century Fox 71 3.081431951 60 $212.00 2011 15
```

## 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.

The screenshot shows the Azure Data Factory pipeline editor interface. The top navigation bar includes tabs for task1q2dataflow1, fortask1source, pipeline1, pipeline2, AzureSqlTable1, pipeline3, and a gear icon. Below the tabs, there's a search bar labeled "Search activities" and a sidebar titled "Activities" with sections for Move & transform, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The main workspace is titled "pipeline3" and contains a single activity node. The "Properties" pane on the right shows the "General" tab selected, with the name set to "pipeline3". The "Annotations" pane at the bottom right has a "New" button. At the bottom of the workspace, there are tabs for Parameters, Variables, Settings, and Output, with the "Parameters" tab currently active.

The screenshot shows the Azure Data Factory pipeline editor interface. At the top, there are tabs for 'task1q2dataflow1', 'fortask1source', 'pipeline1', 'pipeline2', 'AzureSqlTable1', and 'pipeline3'. Below the tabs, the left sidebar has sections for 'Activities' (with 'Search activities' input), 'Move & transform' (including 'Copy data' and 'Data flow'), and various service integrations like 'Azure Data Explorer', 'Azure Function', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning', and 'Power Query'. The main workspace displays a 'Copy data' activity named 'Copy data1'. The activity card includes a preview pane with a red circle icon, a green checkmark, a red X, and a blue refresh button. Below the activity card, the 'General' tab is selected in the properties pane, which contains fields for 'Name' (set to 'Copy data1'), 'Description' (empty), 'Timeout' (set to '0:12:00:00'), 'Retry' (set to '0'), and 'Retry interval (sec)' (set to '30').

**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                     | C                                   |                         |      |                  |       |                   |

[Continue](#) [Cancel](#)

**Set properties**

Name

Linked service \*

Table name  [Edit](#)

Import schema  From connection/store  None

[Advanced](#)

[OK](#) [Back](#) [Cancel](#)

**Set properties**

Name

Linked service \*

File path  /  /  [File](#) [...](#)

First row as header

Import schema  From connection/store  From sample file  None

[Advanced](#)

[OK](#) [Back](#) [Cancel](#)

Validate Validate copy runtime Debug Add trigger

### Properties

- General Related
- Name \* SQL2ADLS
- Description
- Annotations + New

**Mapping**

| Source                | Type        | Destination           | +          | -   |
|-----------------------|-------------|-----------------------|------------|-----|
| Film                  | abc varchar | Film                  | abc String | + - |
| Genre                 | abc varchar | Genre                 | abc String | + - |
| Lead_Studio           | abc varchar | Lead_Studio           | abc String | + - |
| Audience_score_prcnt  | 123 int     | Audience_score_prcnt  | abc String | + - |
| Profitability         | 1.2 float   | Profitability         | abc String | + - |
| Rotten_Tomatoes_prcnt | 123 int     | Rotten_Tomatoes_prcnt | abc String | + - |
| Worldwide_Gross       | abc varchar | Worldwide_Gross       | abc String | + - |
| Year                  | 123 int     | Year                  | abc String | + - |
| FilmCount             | 123 int     | FilmCount             | abc String | + - |

Add dynamic content [Alt+Shift+D]

Validate all Publishing

### 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) | -        |

**Activities**

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDIInsight
- Iteration & conditionals
- Machine Learning
- Power Query

**Mapping**

| Source               | Type        | Destination          |
|----------------------|-------------|----------------------|
| Film                 | abc varchar | Film                 |
| Genre                | abc varchar | Genre                |
| Lead_Studio          | abc varchar | Lead_Studio          |
| Audience_score_prcnt | 123 int     | Audience_score_prcnt |
| Profitability        | 1.2 float   | Profitability        |

**Destinations**

- AzureSqlTable1
- pipeline3
- DelimitedText6
- task1q2dataflow1
- task2q1sol

**Actions**

- Validate
- Validate copy runtime
- Debug
- Add trigger
- Publish
- Cancel

validate all PUBLISH all

task1q2dataflow1 fortask1source pipeline1 pipeline2 AzureSqlTable1 pipeline3

### Activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service

**Trigger now**

**New/Edit**

**Copy data**

task2q1sol

**Actions**

- Validate
- Validate copy runtime
- Debug
- Add trigger

All pipeline runs > SQL2ADLS - Activity runs

### SQL2ADLS

[List](#) [Gantt](#)

[Rerun](#) [Rerun from activity](#) [Rerun from failed activity](#) [Refresh](#) [Update pipeline](#)

Copy data ✓

[+](#) [-](#) [\[ \]](#) [\[ \]](#)

#### Activity runs

Pipeline run ID d965cdfc-d935-4e2f-aca8-6d62fc6cf648

All status ▾

Showing 1 - 1 items

| Activity name | Activity type | Run start               | Duration | Status                                         | Error | Log | Integration run |
|---------------|---------------|-------------------------|----------|------------------------------------------------|-------|-----|-----------------|
| Copy data1    | Copy data     | Oct 8, 2022, 3:27:42 pm | 00:00:09 | <span style="color: green;">✓ Succeeded</span> |       |     | AutoResolver    |

Home > Storage accounts > finalprojadlsstorage | Containers >

**solution** Container

[Search](#) [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, 3:27:50 PM | Hot (Inferred) |                | Block blob | 6.38 KB | Available   |
| GenreCount          | 10/8/2022, 3:14:40 PM | Hot (Inferred) |                | Block blob | 72 B    | Available   |

| A  | B            | C       | D           | E        | F            | G               | H         | I    | J  | K         | L | M | N | O |
|----|--------------|---------|-------------|----------|--------------|-----------------|-----------|------|----|-----------|---|---|---|---|
| 1  | Film         | Genre   | Lead_Studio | Audience | Profitabilit | Rotten_Tomatoes | Worldwide | Year |    | FilmCount |   |   |   |   |
| 2  | Zack and N   | Romance | The Weins   | 70       | 1.747542     | 64              | \$41.94   | 2008 | 15 |           |   |   |   |   |
| 3  | Waitress     | Romance | Independ    | 67       | 11.08974     | 89              | \$22.18   | 2007 | 15 |           |   |   |   |   |
| 4  | Waiting Fc   | Romance | Independ    | 53       | 0.005        | 6               | \$0.03    | 2011 | 15 |           |   |   |   |   |
| 5  | Tyler Perry  | Romance | Independ    | 47       | 3.724192     | 46              | \$55.86   | 2007 | 15 |           |   |   |   |   |
| 6  | Twilight: B  | Romance | Independ    | 68       | 6.383364     | 26              | \$702.17  | 2011 | 15 |           |   |   |   |   |
| 7  | Twilight     | Romance | Summit      | 82       | 10.18003     | 49              | \$376.66  | 2008 | 15 |           |   |   |   |   |
| 8  | Something    | Romance | Independ    | 48       | 1.719514     | 15              | \$60.18   | 2011 | 15 |           |   |   |   |   |
| 9  | P.S. I Love  | Romance | Independ    | 82       | 5.103117     | 21              | \$153.09  | 2007 | 15 |           |   |   |   |   |
| 10 | One Day      | Romance | Independ    | 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 | Independ    | 84       | 0.652603     | 54              | \$29.37   | 2007 | 15 |           |   |   |   |   |
| 17 | The Curious  | 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 Twilight | Drama   | Summit      | 78       | 14.1964      | 27              | \$709.82  | 2009 | 13 |           |   |   |   |   |
| 20 | The Time     | Drama   | Paramoun    | 65       | 2.598205     | 38              | \$101.33  | 2009 | 13 |           |   |   |   |   |
| 21 | The Duke     | Drama   | Paramoun    | 68       | 3.20785      | 60              | \$43.31   | 2008 | 13 |           |   |   |   |   |
| 22 | Remember     | Drama   | Summit      | 70       | 3.49125      | 28              | \$55.86   | 2010 | 13 |           |   |   |   |   |
| 23 | Rachel Get   | Drama   | Independ    | 61       | 1.384167     | 85              | \$16.61   | 2008 | 13 |           |   |   |   |   |
| 24 | Not Easily   | Drama   | Independ    | 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.

The screenshot shows the Azure Data Factory interface. On the left, the 'Activities' pane is open, showing various options like Move & transform, Azure Data Explorer, and Databricks. Under Databricks, the 'Notebook' option is selected and highlighted with a red circle (Step 1). In the main workspace, a 'Copy data' activity is connected to a 'Notebook' activity (Step 2). Below the notebook activity, there is a 'Databricks linked service' dropdown with a 'New...' button highlighted with a red circle (Step 3).

This screenshot shows the 'New linked service' dialog for Azure Databricks. It includes fields for Name (AzureDatabricks1), Description, Connect via integration runtime (AutoResolveIntegrationRuntime), Account selection method (From Azure subscription), Azure subscription (Select all), Databricks workspace (Databricks workspace URL), Select cluster (New job cluster), and Authentication type. At the bottom, there is a 'Create' button and a 'Test connection' link.

This screenshot shows the Microsoft Azure Databricks portal. It displays the 'Create a cluster' wizard, which has completed the 'Import data' step. The main configuration page for 'Vaibhav Pandey's Cluster' is shown, including settings for Access mode (Single user access), Performance (Databricks runtime version: Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1), Node type: Standard\_DS3\_v2, Terminate after 30 minutes of inactivity), and Tags. A summary table on the right provides details about the cluster configuration. At the bottom, there is a 'Create Cluster' button.

Microsoft Azure databricks

Search CTRL + P

Vaibhav Pandey's Cluster UI Preview Provide feedback

Free trial ends in 13 days. Upgrade to Premium in Azure Portal

More ... Terminate Edit

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

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

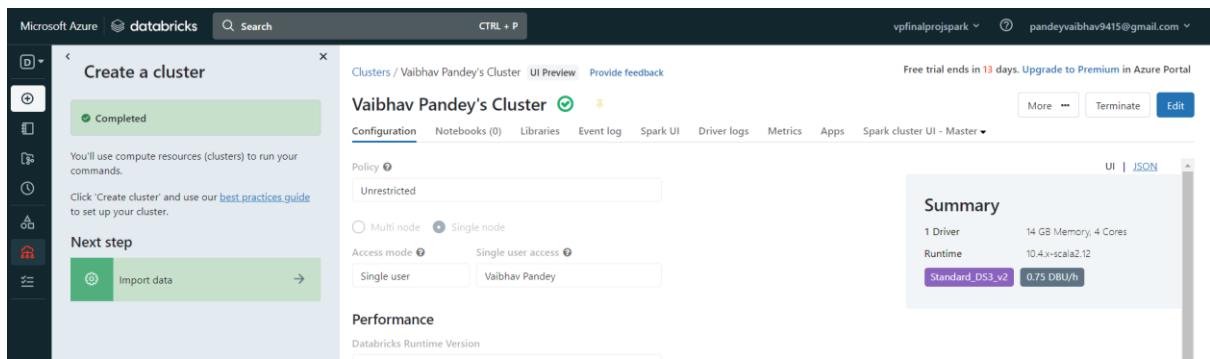
Performance Databricks Runtime Version

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

Copy data task2q1sol

Notebook

New linked service

Azure Databricks Learn more

Databricks workspace vpfinalprospark

Select cluster

New job cluster Existing interactive cluster Existing instance pool

Databrick Workspace URL https://adb-3611621693894286.azuredatabricks.net

Authentication type Access Token

Access token

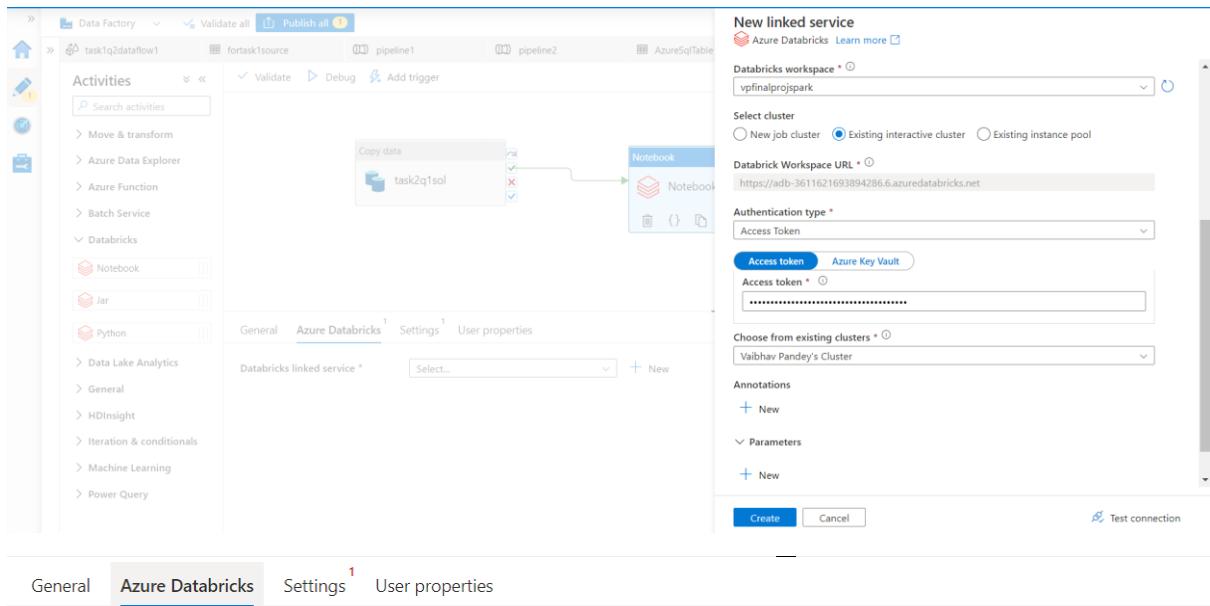
Access token \* (redacted)

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

Vaibhav Pandey's Cluster UI Preview Provide feedback

Free trial ends in 13 days. Upgrade to Premium in Azure Portal

More ... Terminate Edit

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

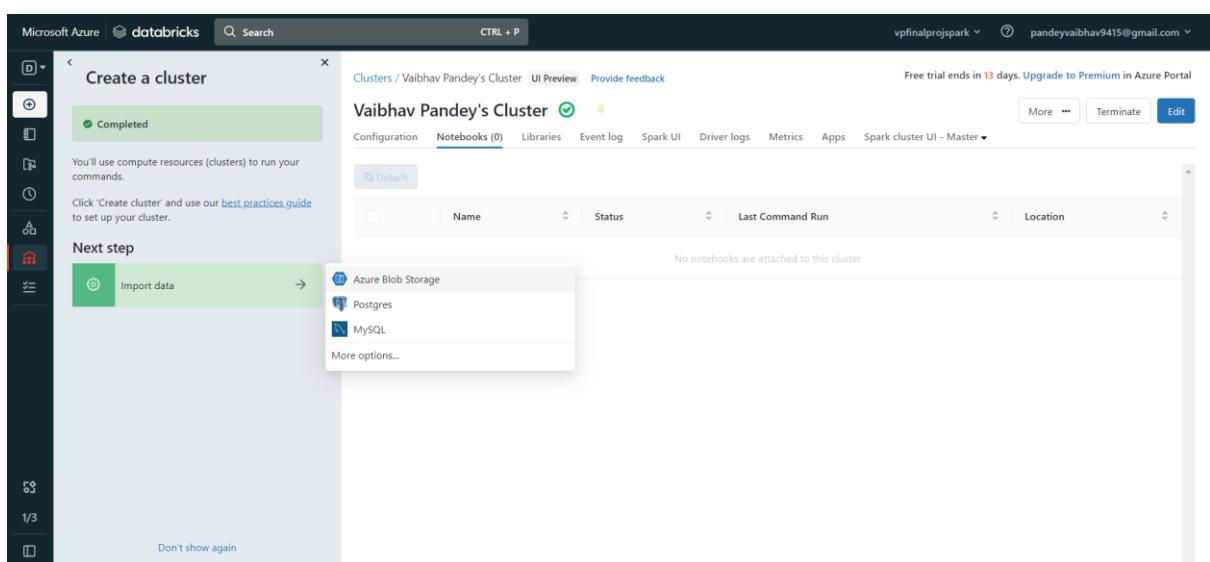
Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark cluster UI - Master

Detach

| Name                                      | Status | Last Command Run | Location |
|-------------------------------------------|--------|------------------|----------|
| No notebooks are attached to this cluster |        |                  |          |

Azure Blob Storage Postgres MySQL More options...

1/3 Don't show again



**Microsoft Azure databricks** Search CTRL + P vpfinalprospark 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

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,"/btQwJNc7ASBqGLW6DDMpR2PBqj3SnC/28LhyqqDFbVUEVN
4 JupIMlOM7AaqRYSMyFvI0kgKp+ND+ASTNYUBwg==")

Command took 0.84 seconds -- by pandeyvaibhav9415@gmail.com at 10/8/2022, 11:04:20 AM on Vaibhav Pandey's Cluster
```

**Home > Storage accounts > finalprojadlsstorage**

**finalprojadlsstorage | 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   |
| sq2adlsfinaldest | 10/8/2022, 12:06:32 AM | Private             | Available   |
| task2q2          | 10/8/2022, 3:35:54 PM  | Private             | Available   |

**task2q2**

Container

Upload Completed for part.csv 6.38 KB | 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

```
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
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
```

```

1 df=spark.read.option('header',True).option('inferSchema',True).csv("abfss://task2q@finalprojadlsstorage.dfs.core.windows.net/movies/part.csv")
2
3 df.show()
4

```

▶ (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      | Independent      | 67  11.0897415       | 89            | \$22.18  2007         | 15              |      |           |
| Waiting For Forever Romance                       | Independent      | Independent      | 53  0.005            | 6             | \$0.03  2011          | 15              |      |           |
| Tyler Perry's Why... Romance                      | Independent      | Independent      | 47  3.7241924        | 46            | \$55.86  2007         | 15              |      |           |
| Twilight: Breakin... Romance                      | Independent      | Independent      | 68  6.383363636      | 26            | \$702.17  2011        | 15              |      |           |
| Twilight Romance                                  | Summit           | Summit           | 82  10.18002703      | 49            | \$376.66  2008        | 15              |      |           |
| Something Borrowed Romance                        | Independent      | Independent      | 48  1.719514286      | 15            | \$60.18  2011         | 15              |      |           |
| P.S. I Love You Romance                           | Independent      | Independent      | 82  5.183116833      | 21            | \$153.09  2007        | 15              |      |           |
| One Day Romance                                   | Independent      | Independent      | 54  3.682733333      | 37            | \$55.24  2011         | 15              |      |           |
| New Year's Eve Romance                            | Warner Bros.     | Warner Bros.     | 48  2.536428571      | 8             | \$142.84  2011        | 15              |      |           |
| Music and Lyrics Romance                          | Warner Bros.     | Warner Bros.     | 70  3.64741085       | 63            | \$145.98  2007        | 15              |      |           |
| Monte Carlo Romance                               | 20th Century Fox | 20th Century Fox | 50  1.9832           | 38            | \$39.66  2011         | 15              |      |           |
| Midnight in Paris Romance                         | Sony             | Sony             | 84  8.744705882      | 93            | \$148.66  2011        | 15              |      |           |
| Jane Eyre Romance                                 | Universal        | Universal        | 77  0.0              | 85            | \$30.15  2011         | 15              |      |           |
| Across the Universe Romance                       | Independent      | Independent      | 84  0.652603178      | 54            | \$29.37  2007         | 15              |      |           |
| The Curious Case ... Fantasy                      | Warner Bros.     | Warner Bros.     | 81  1.78394375       | 73            | \$285.43  2008        | 1               |      |           |
| Water for Elephants  Drama                        | 20th Century Fox | 20th Century Fox | 72  3.081421053      | 69            | \$117.89  2011        | 13              |      |           |
| The Twilight Saga...  Drama                       | Summit           | Summit           | 78  14.19641         | 27            | \$709.82  2009        | 13              |      |           |

Command took 4.41 seconds -- by pandeyvatbhav9415@gmail.com at 10/8/2022, 3:49:45 PM on Vaibhav Pandey's Cluster

```
1 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 pandeyvatbhav9415@gmail.com at 10/8/2022, 3:49:54 PM on Vaibhav Pandey's Cluster

Cmd 7

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

▶ (1) Spark Jobs

▶ df: 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  2018  | 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               |      |           |                |
| Gnomo and Juliet Animation   | Disney               | 52  5.387972222 | 56                   | \$193.97  2011 | 4                     | 2               |      |           |                |
| Gnomo 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  2.12944167  | 30                   | \$127.77  2007 | 43                    | 6               |      |           |                |
| Sex and the City 2  Comedy   | Warner Bros.         | 49  2.8835      | 15                   | \$288.35  2018 | 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 Mel  Comedy       | Fox                  | 77  3.746781818 | 63                   | \$206.07  2008 | 43                    | 6               |      |           |                |

```
1 %sql
2 select *,RANK () OVER(ORDER BY Genre) as Rankedby_Genre from movies;
```

▶ (1) Spark Jobs

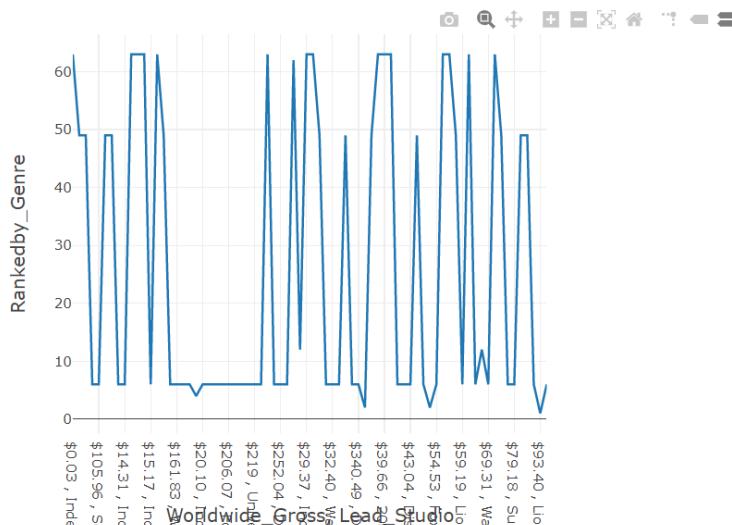
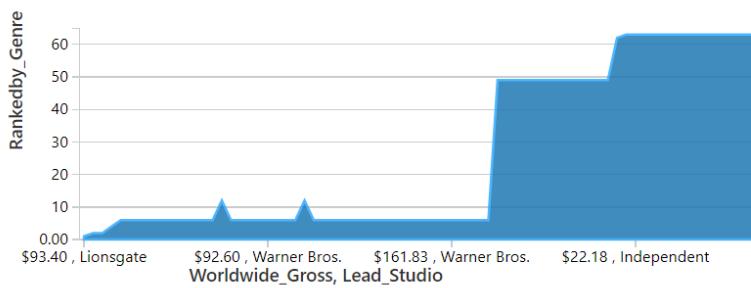
▶ df: pyspark.sql.dataframe.DataFrame = [Film: string, Genre: string ... 8 more fields]

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 | Gnomo and Juliet                   | Animation | Disney                | 52                   | 5.387972222   | 56                    | \$193.97        | 2011 | 4         |
| 5 | Gnomo 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.

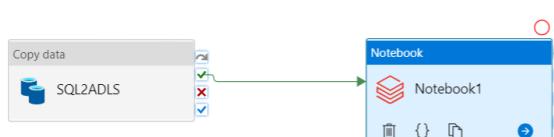
Chart Data Profile



|   | 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



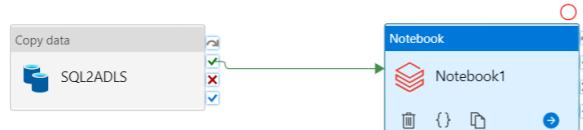
General Azure Databricks Settings User properties

Databricks linked service \*

AzureDatabricks1

Test connection Edit New

Connection successful



The screenshot shows the 'Settings' tab of a Databricks notebook configuration. It includes fields for 'Notebook path' (set to '/Users/pandeyvaibhav9415@gmail.com/da'), 'Base parameters', and 'Append libraries'. On the right, a 'Publish all' dialog is open, showing pending changes for pipelines (SQL2ADLS) and notebooks (Notebook1). The 'Publish' button is highlighted.

The screenshot shows the 'Pipeline runs' history for the 'SQL2ADLS' pipeline. It displays a completed run with two activities: 'Notebook1' (Notebook type) and 'SQL2ADLS' (Copy data type), both marked as 'Succeeded'. The run started on Oct 8, 2022, at 4:11:10 pm.

The screenshot shows the Azure Storage Explorer interface for a container named 'solution'. It lists blobs: 'dbo.movieoutput.csv' (modified 10/8/2022, 4:01:59 PM) and 'GenreCount' (modified 10/8/2022, 3:14:40 PM). The 'dbo.movieoutput.csv' blob is highlighted with a yellow box.

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

**Set properties**

**Name**: DelimitedText8

**Linked service**: AzureDataLakeStorage1

**File path**: task2q2/movies/dbo.movieoutput.csv

**First row as header**:

**Import schema**:  From connection/store  From sample file  None

**Advanced**

**OK** **Back** **Cancel**

**vpfinalprojsynapse | SQL pools**

**Resuming dedicated SQL pool**: Resuming dedicated SQL pool 'vpfinalprojsqlpool' in workspace 'vpfinalprojsynapse'.

| Name               | Type       | Status | Size   |
|--------------------|------------|--------|--------|
| Built-in           | Serverless | N/A    | Auto   |
| vpfinalprojsqlpool | Dedicated  | Paused | DW100c |

**New linked service**

**Azure Synapse Analytics**: Learn more

**vpfinalprojsynapse (Synapse workspace)**

**Database name**: vpfinalprojsqlpool

**SQL pool**: vpfinalprojsqlpool

**Authentication type**: SQL authentication

**User name**: sqladminfinal

**Password**: Azure Key Vault

**Additional connection properties**

**Create** **Cancel** **Test connection**

**Publish all**

You are about to publish all pending changes to the live environment. [Learn more](#)

**Pending changes (4)**

| NAME                        | CHANGE   | EXISTING |
|-----------------------------|----------|----------|
| SQL2ADLS                    | (Edited) | SQL2ADLS |
| ADLS2synapse                | (New)    | -        |
| DelimitedText8              | (New)    | -        |
| AzureSynapseAnalyticsTab... | (New)    | -        |

**General**

Name: ADLStoSynapse

Description:

Timeout: 0:12:00:00

Retry: 0

**Source**

Copy data

ADLStoSynapse

**Mapping**

**Settings**

**User properties**

**Activity runs**

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 | Integration r... |
|---------------|---------------|-------------------------|----------|-----------|-------|-----|------------------|
| ADLStoSynapse | Copy data     | Oct 8, 2022, 6:38:31 pm | 0:00:29  | Succeeded |       |     | AutoResolver     |

**SQLQuery3.sql - vp...sqladminuser (129)\***

```
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;
```

**SQLQuery2.sql - vp...sqladminfinal (73)\***

**SQLQuery1.sql - vp...pdb (sqladmin (68))\***

100 %

Results Messages

| Film                               | Genre     | Lead_Studio           | Audience_score_prcnt | Profitability | Rotten_Tomatoes_prcnt | Worldwide_Gross | Year | FilmCount | RankedBy_Genre |
|------------------------------------|-----------|-----------------------|----------------------|---------------|-----------------------|-----------------|------|-----------|----------------|
| Tangled                            | Animation | Disney                | 88                   | 1.365692308   | 89                    | \$355.01        | 2010 | 4         | 2              |
| Youth in Revolt                    | Comedy    | The Weinstein Company | 52                   | 1.09          | 68                    | \$19.62         | 2010 | 43        | 6              |
| WALL-E                             | Animation | Disney                | 89                   | 2.896019067   | 96                    | \$521.28        | 2008 | 4         | 2              |
| Killers                            | Action    | Lionsgate             | 45                   | 1.245333333   | 11                    | \$93.40         | 2010 | 1         | 1              |
| You Will Meet a Tall Dark Stranger | Comedy    | Independent           | 35                   | 1.211818182   | 43                    | \$26.66         | 2010 | 43        | 6              |
| 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              |
| Valentine's Day                    | Comedy    | Warner Bros.          | 54                   | 4.184038462   | 17                    | \$217.57        | 2010 | 43        | 6              |
| The Invention of Lying             | Comedy    | Warner Bros.          | 47                   | 1.751351351   | 56                    | \$32.40         | 2009 | 43        | 6              |
| What Happens in Vegas              | Comedy    | Fox                   | 72                   | 6.267647029   | 28                    | \$219.37        | 2008 | 43        | 6              |
| When in Rome                       | Comedy    | Disney                | 44                   | 0             | 15                    | \$43.04         | 2010 | 43        | 6              |
| The Heartbreak Kid                 | Comedy    | Paramount             | 41                   | 2.12944167    | 30                    | \$127.77        | 2007 | 43        | 6              |
| The Ugly Truth                     | Comedy    | Independent           | 68                   | 5.402631579   | 14                    | \$205.30        | 2009 | 43        | 6              |
| The Proposal                       | Comedy    | Disney                | 74                   | 7.8675        | 43                    | \$314.70        | 2009 | 43        | 6              |
| Sex and the City Two               | Comedy    | Warner Bros.          | 49                   | 2.8835        | 15                    | \$288.35        | 2010 | 43        | 6              |
| Penelope                           | Comedy    | Summit                | 74                   | 1.382799733   | 52                    | \$20.74         | 2008 | 43        | 6              |
| Shrek Out of Africa                | Comedy    | Paramount             | 60                   | 2.4406        | 67                    | \$10.91         | 2010 | 43        | 6              |

Query executed successfully.

vpfinalprojsynapse.sqlazur... sqladminuser (129) vpfinalprojsqlpool 00:00:00 | 77 rows