

# 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:

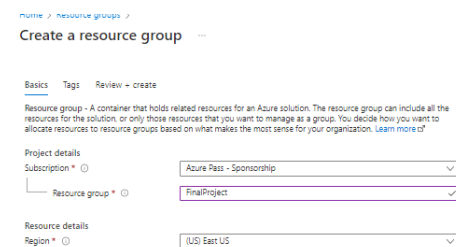
As a Data Engineer, you've been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will be collecting the data from SQL Database using ADF and the transformed data will be the source for databricks to run complex 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.



The screenshot shows the 'Create a resource group' page in the Azure portal. The page has a breadcrumb 'Home > Resource groups >' and a title 'Create a resource group'. Below the title are tabs for 'Basics', 'Tags', and 'Review + create', with 'Basics' being the active tab. A descriptive paragraph explains that a resource group is a container for related resources. The 'Project details' section contains two dropdown menus: 'Subscription' (set to 'Azure Pass - Sponsorship') and 'Resource group' (set to 'FinalProject'). The 'Resource details' section contains a 'Region' dropdown menu set to '(US) East US'. At the bottom, there are three buttons: 'Review + create', '< Previous', and 'Next > Tags'.

## Resource Group → FinalProject

Home > Resource groups > FinalProject

Resource groups

Filter for any field...

Name 1:

FinalProject

Subscription (copy) : Azure Pass - Sponsorship

Access control (IAM)

Tags (edit) : Click here to add tags

Deployments : No deployments

Location : East US

Resources Recommendations

Showing 0 to 0 of 0 records. ☐ Show hidden types

No resources match your filters. Try changing or clearing your filters.

Create resources Clear filters

Learn more

## • Create a Storage account.

### a) Blob Storage

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review + create

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription \* Azure Pass - Sponsorship

Resource group \* FinalProject

Instance details

If you need to create a legacy storage account type, please click [here](#).

Storage account name \* rvblobfinalproject

Region \* (US) East US

Performance \* ☒ Standard: Recommended for most scenarios (general-purpose v2 account) ☐ Premium: Recommended for scenarios that require low latency.

Redundancy \* Locally-redundant storage (LRS)

Review + create < Previous Next: Advanced >

## Blob storage → rvblobfinalproject

Home > Storage accounts > rvblobfinalproject

Storage account

Filter for any field...

rvblobfinalproject

Subscription (copy) : Azure Pass - Sponsorship

Access control (IAM)

Tags (edit) : Click here to add tags

Performance : Standard

Replication : Locally-redundant storage (LRS)

Account kind : StorageV2 (general purpose v2)

Provisioning state : Succeeded

Created : 4/5/2022, 3:14:03 PM

Properties Monitoring Capabilities (7) Recommendations Tutorials Developer Tools

Blob service

Hierarchical namespace : Disabled

Default access tier : Hot

Blob public access : Enabled

Blob soft delete : Enabled (7 days)

Container soft delete : Enabled (7 days)

Versioning : Disabled

Change feed : Disabled

NFS v3 : Disabled

Allow cross-tenant replication : Enabled

File service

Large file share : Disabled

Active Directory : Not configured

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

Networking

Allow access from : All networks

Number of private endpoint connections : 0

Network routing : Microsoft network routing

Access for trusted Microsoft services : Yes

## b) Azure Data Lake Storage → rvadlsfinalproject

The screenshot displays the Microsoft Azure portal interface for the 'rvadlsfinalproject' Storage account. The left sidebar shows navigation options like Overview, Activity log, Tags, and Data storage. The main content area is divided into 'Essentials' and 'Properties' sections. The 'Essentials' section provides key information: Resource group (FinalProject), Location (East US), Primary/Secondary Location (Primary: East US, Secondary: West US), Subscription (Azure Pass - Sponsorship), Subscription ID (d5944b87-9fd2-4636-bf6f-44aa5575d1cc), and Disk state (Primary: Available, Secondary: Available). The 'Properties' section includes 'Data Lake Storage' settings (Hierarchical namespace: Disabled, Default access tier: Hot, Blob public access: Enabled, Blob soft delete: Enabled (7 days), Container soft delete: Enabled (7 days), Versioning: Disabled, Change feed: Disabled, NFS v3: Disabled, SFTP: Disabled) and 'File service' settings (Large file share: Disabled, Active Directory: Not configured, Soft delete: Enabled (7 days)). A 'Security' section on the right shows settings for secure transfer, storage account key access, minimum TLS version, and infrastructure encryption. A 'Networking' section shows settings for allow access from, number of private endpoint connections, network routing, and access for trusted Microsoft services.

## • Create an Azure SQL Database.

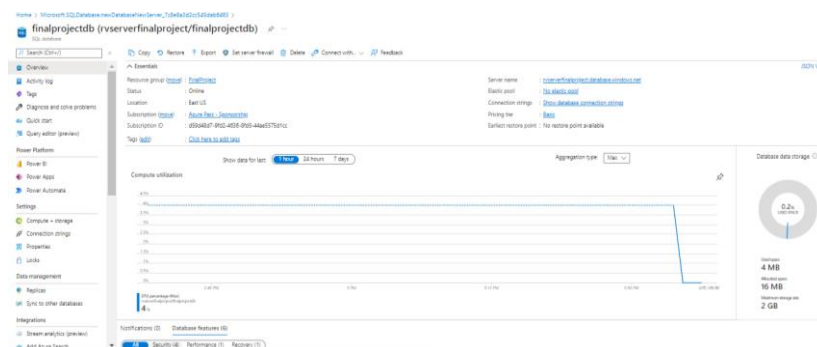
### Server

The screenshot shows the 'Create SQL Database Server' form in the Microsoft Azure portal. The form is titled 'Create SQL Database Server' and includes a 'Server details' section. The 'Server name' field is set to 'rvadlsfinalproject' and the 'Location' is set to 'East US'. The 'Authentication method' section has three options: 'Use SQL authentication' (selected), 'Use only Azure Active Directory (Azure AD) authentication', and 'Use both SQL and Azure AD authentication'. The 'Server admin login' field is set to 'sqladminuser', and the 'Password' and 'Confirm password' fields are filled with asterisks. The 'OK' button is visible at the bottom.

### Sql Database

The screenshot shows the 'Create SQL Database' form in the Microsoft Azure portal. The form is titled 'Create SQL Database' and includes a 'Database details' section. The 'Subscription' is set to 'Azure Pass - Sponsorship' and the 'Resource group' is set to 'FinalProject'. The 'Database name' field is set to 'finalprojectdb'. The 'Server' is set to '(new) rvadlsfinalproject (East US)'. The 'Compute + storage' section has three options: 'Basic' (selected), '2 GB storage', and 'Configure database'. The 'Backup storage redundancy' section has three options: 'Locally-redundant backup storage' (selected), 'Zone-redundant backup storage', and 'Geo-redundant backup storage'. The 'Review + create' button is visible at the bottom.

## Sql Database --> finalprojectdb



## • Create a data factory.

Home > Data factories > Create Data Factory

Changes on this step may reset later selections you have made. Review all options prior to deployment.

Basics Git configuration 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 (🔍) Azure Pass - Sponsorship

Resource group (🔍) FinalProject

Instance details

Name (🔍) rvfinalprojectdf

Region (🔍) East US

Version (🔍) V2 (Recommended)

Review & create < Previous Next: Get configuration >

Data Factory → rvfinalprojectdf

Home > Microsoft.DataFactory-20220405154754 > rvfinalprojectdf

Data factory (V2)

Search (Ctrl+F) Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Networking

Managed identities

Properties

Locks

Getting started

Quick start

Monitoring

Alerts

Metrics

Diagnostic settings

Logs

Automation

Tasks (preview)

Support + troubleshooting

Essentials

Resource group (🔍) FinalProject

Status Succeeded

Location East US

Subscription (🔍) Azure Pass - Sponsorship

Subscription ID d59d4bd7-9f02-4d56-bfd5-44ae5575d1cc

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

PipelineRuns

ActivityRuns

TriggerRuns

Integration Runtime CPU

## • Configure Databricks cluster

### Creating Azure Databricks

Home > Azure Databricks > 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 (🔍) Azure Pass - Sponsorship

Resource group (🔍) FinalProject

Instance Details

Workspace name (🔍) rvfinalprojectdatabricks

Region (🔍) East US

Pricing tier (🔍) Trial (Premium - 14-Days Free DBU)

Review & create < Previous Next: Networking >

## Databricks → rvfinalprojectdatabricks

The screenshot shows the Azure Databricks portal for the workspace 'rvfinalprojectdatabricks'. The left sidebar contains navigation links: Overview, Activity log, Access control (IAM), Tags, Settings, Virtual Network Peering, Encryption, Properties, Locks, Automation, Tasks (preview), Export template, Support + troubleshooting, and New Support Request. The main area displays the workspace's status as 'Active' and provides links to 'Launch Workspace' and 'Upgrade to Premium'. Below this, there are several tiles for 'Documentation', 'Getting Started', 'Import Data from File', 'Import Data from Azure Storage', 'Notebook', 'Admin Guide', and 'Link Azure ML workspace'.

## Creating Cluster → rvfinalprojectcluster

This screenshot shows the 'Create a cluster' dialog in the Azure Databricks portal. The dialog is titled 'Create Cluster' and includes a 'New Cluster' section with a 'Cancel' button and a 'Create Cluster' button. The cluster configuration is as follows: Cluster name is 'rvfinalprojectcluster'; Cluster mode is 'Single Node'; Databricks runtime version is 'Runtime: 9.1 LTS (Scala 2.12, Spark 3.1.2)'; Autopilot options are checked with 'Terminate after 120 minutes of inactivity'; and Node type is 'Standard\_DS3\_v2' with '14 GB Memory, 4 Cores'. A promotional discount banner for 50% off is visible. The left sidebar shows the 'Create a cluster' dialog with a 'Don't show again' link.

This screenshot shows the 'rvfinalprojectcluster' cluster page in the Azure Databricks portal. The cluster is in a 'Completed' state. The left sidebar shows the 'Create a cluster' dialog with a 'Next step' section containing an 'Ingest data' button. The main area displays the cluster's configuration: Policy is 'Unrestricted'; Cluster mode is 'Single Node'; Databricks Runtime Version is '9.1 LTS (includes Apache Spark 3.1.2, Scala 2.12)'; Autopilot options are checked with 'Terminate after 120 minutes of inactivity'; and Node type is 'Standard\_DS3\_v2' with '14 GB Memory, 4 Cores'. The top right corner indicates 'Free trial ends in 14 days. Upgrade to Premium in Azure Portal'.

## ● Create Synapse analytics Data Warehouse.

Microsoft Azure

Home > Azure Synapse Analytics >

### Create Synapse workspace

Learn 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 to organize and manage all of your resources.

Subscription: Azure Pass - Sponsorship

Resource group: rfvfinalproject

Managed resource group: rfvfinalproject

**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: rfvfinalprojectworkspace

Region: East US

Select Data Lake Storage Gen2: From subscription

Account name: rfvfinalprojectworkspace

File system name: rfvfinalprojectworkspace

Review + create

Azure Synapse Analytics

Default Directory

+ Create + Manage view

Filter for any field...

Name

- raevenasynapse
- raevenaproject
- rfvfinalprojectworkspace

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

- Encryption
- Networking
- Identity
- Private endpoint connections
- Approved Azure AD tenants
- Azure SQL Auditing
- Microsoft Defender for Cloud

Monitoring

Essentials

Resource group: rfvfinalproject

Status: Succeeded

Location: East US

Subscription: Azure Pass - Sponsorship

Subscription ID: d59d48b7-...

Managed virtual network: No

Managed identity object: e68e2945-280e-4d52-a269-3d45167e68d0

Workspace web URL: https://web.azure.synapse.net/workspaces/%7b%22subscription%22%3A%22d59d48b7-9f92-4d63-...

Tags

Getting started

Open Synapse Studio

Read documentation

Analytics pools

Name	Type	Size
Loading...		

## SQLPOOL

Home > Microsoft Azure Synapse SQL Pool On Existing Workspace\_25342148b9824 > rfvfinalprojectworkspace >

### sqlpoolforproject (rfvfinalprojectworkspace/sqlpoolforproject)

Dedicated SQL pool

Search (Ctrl+F)

Overview

Activity log

Access control (IAM)

Tags

Settings

- Workload management
- Maintenance schedule
- Geo-backup policy
- Connection strings
- Properties
- Locks

Security

- Auditing
- Data Discovery & Classification
- Dynamic Data Masking
- Microsoft Defender for Cloud
- Transparent data encryption

Common Tasks

- Open in Visual Studio

Monitoring

- Query activity

Essentials

Resource group: rfvfinalproject

Status: Online

Location: East US

Subscription: Azure Pass - Sponsorship

Subscription ID: d59d48b7-...

Tags

Notifications (0)

Features (4)

Tasks (2)

Alerts (0)

Recommendations (0)

Info (0)

There are no notifications to display.

Security (3)

Recovery (1)

Transparent data encryption

Auditing

Microsoft Defender for SQL

Geo-backup

Query in Visual Studio

Monitor

DWU usage

Active and queued queries

- 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

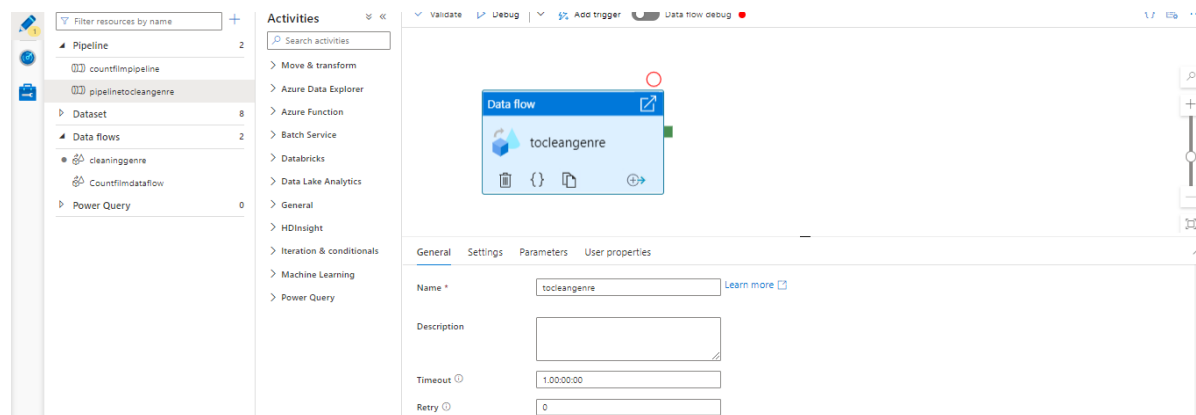
### Task 1: Create a dataflow with the following requirement:

1. Create a data stream named CleaningGenreRomance and perform data cleansing on the Genrecolumn 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.)

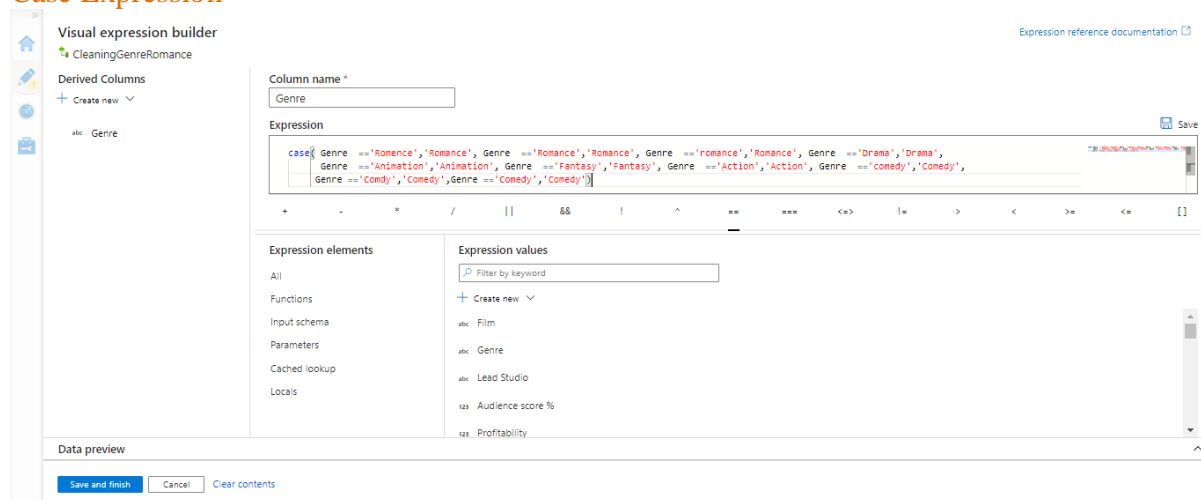
### Dataflow cleaninggenre

The screenshot displays the Microsoft Azure Data Factory console. On the left, the 'Factory Resources' pane shows a list of resources including AzureSqlTable3 through AzureSqlTable9, DelimitedText1 through DelimitedText9, and a 'movies' dataset. Under the 'Data flows' section, 'cleaninggenre' is selected. The main workspace shows a dataflow diagram with three activities: 'fromadls' (Import data from DelimitedText5), 'CleaningGenreRomance' (Creating/updating the columns Film, Genre, Lead Studio, Audience score %, Profitability, Rotten Tomatoes %, Worldwide), and 'toadlsoutput' (Export data to DelimitedText3). Below the diagram, the 'Sink' settings for the 'CleaningGenreRomance' activity are visible, showing 'Output stream name' as 'tosqldatabase', 'Incoming stream' as 'CleaningGenreRomance', 'Sink type' as 'Dataset', 'Dataset' as 'AzureSqlTable6', and 'Options' with 'Allow schema drift' checked.

## Datapipeline pipelinetocleangenre



## Case Expression



```
case( Genre == 'Romence', 'Romance', Genre == 'Romance', 'Romance', Genre == 'romance', 'Romance', Genre == 'romance', 'Romance', Genre == 'Drama', 'Drama', Genre == 'Animation', 'Animation', Genre == 'Fantasy', 'Fantasy', Genre == 'Action', 'Action', Genre == 'comedy', 'Comedy', Genre == 'Comdy', 'Comedy', Genre == 'Comedy', 'Comedy' )
```



## Output in ADLS:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
9	Wallace	Romance	Indepand	53	0.005	9	\$22.18	2007														
10	Valentine	Comedy	Warner E	54	4.184	17	\$217.57	2010														
11	Tyler Perry	Romance	Indepand	47	3.1242	46	\$55.86	2007														
12	Twilight	Romance	Indepand	69	6.3524	26	\$702.17	2011														
13	Twilight	Romance	Summit	82	10.18	49	\$376.66	2008														
14	The Ugly Comedy	Indepand	68	5.4026	14	\$205.20	2009															
15	The Ugly Comedy	Indepand	70	14.196	27	\$703.62	2009															
16	The Time Drama	Paranorm	65	2.5962	38	\$101.33	2009															
17	The Time Drama	Comedy	Disney	14	1.0615	43	\$314.70	2009														
18	The Time Drama	Comedy	Warner E	47	1.7514	56	\$32.40	2009														
19	The Time Drama	Comedy	Paranorm	41	2.1634	30	\$127.77	2007														
20	The Time Drama	Comedy	Paranorm	69	3.2073	60	\$43.31	2008														
21	The Time Drama	Comedy	Warner E	81	1.7839	73	\$285.43	2008														
22	The Time Drama	Comedy	CBS	47	2.2026	20	\$77.09	2010														
23	The Time Drama	Comedy	Disney	68	1.5657	63	\$355.01	2010														
24	The Time Drama	Comedy	Indepand	48	1.7195	15	\$60.18	2011														
25	The Time Drama	Comedy	Paranorm	60	2.4405	57	\$40.61	2010														
26	The Time Drama	Comedy	Warner E	49	2.6835	15	\$288.35	2010														
27	The Time Drama	Comedy	Warner E	43	2.6835	15	\$288.35	2010														
28	The Time Drama	Comedy	Warner E	81	1.2218	49	\$495.25	2008														
29	The Time Drama	Comedy	Summit	70	3.4913	28	\$55.86	2010														
30	The Time Drama	Comedy	Indepand	61	1.3842	85	\$16.61	2008														
31	The Time Drama	Comedy	Summit	14	1.3828	52	\$10.14	2008														
32	The Time Drama	Comedy	Indepand	82	5.1031	21	\$153.09	2007														
33	The Time Drama	Comedy	New Line	47	2.071	15	\$20.71	2008														
34	The Time Drama	Comedy	Indepand	43	0	14	\$21.37	2010														
35	The Time Drama	Comedy	Indepand	54	3.6627	37	\$55.24	2011														
36	The Time Drama	Comedy	Indepand	66	2.14	34	\$10.70	2009														
37	The Time Drama	Comedy	Warner E	64	3.3012	39	\$32.60	2007														
38	The Time Drama	Comedy	Soap	67	3.3927	73	\$33.33	2008														
39	The Time Drama	Comedy	Warner E	48	2.5564	8	\$142.04	2011														
40	The Time Drama	Comedy	The War	84	0.8258	83	\$9.26	2011														
41	The Time Drama	Comedy	Warner E	70	3.6474	63	\$145.90	2007														
42	The Time Drama	Comedy	20th Cent	50	1.3832	38	\$23.66	2011														
43	The Time Drama	Comedy	Indepand	70	0.2523	78	\$15.17	2008														
44	The Time Drama	Comedy	Soap	84	8.7447	93	\$149.66	2011														
45	The Time Drama	Comedy	Fox	77	1.7463	63	\$206.07	2008														
46	The Time Drama	Comedy	Univers	76	3.2345	53	\$609.47	2008														

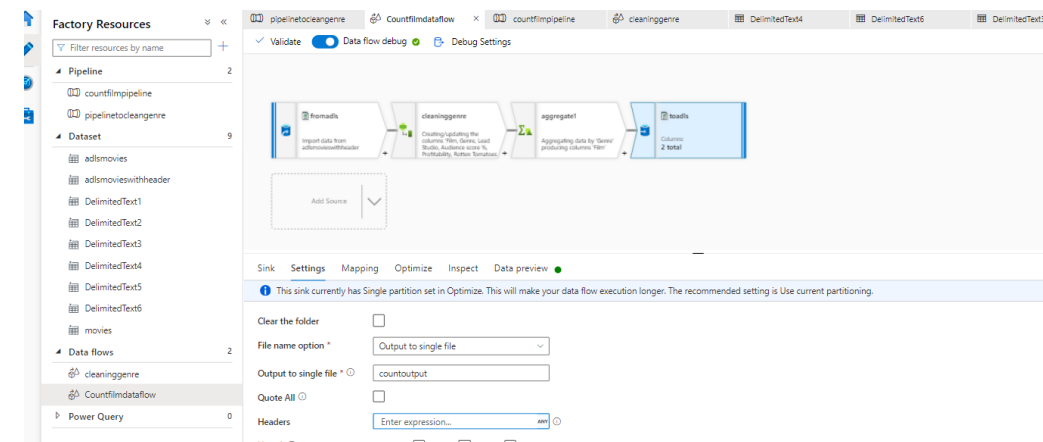
	A	B	C	D	E	F	G	H	I	J
46	Marley's Comedy	Comedy	Fox	77	3.7458	63	\$206.07	2008		
47	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
48	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
49	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
50	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
51	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
52	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
53	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
54	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
55	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
56	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
57	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
58	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
59	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
60	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
61	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
62	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
63	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
64	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
65	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
66	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
67	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
68	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
69	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
70	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
71	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
72	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
73	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
74	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
75	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
76	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
77	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
78	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
79	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
80	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
81	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
82	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
83	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			
84	Mamma's Comedy	Univers	76	3.2345	53	\$609.47	2008			

## Output in Sql Database :

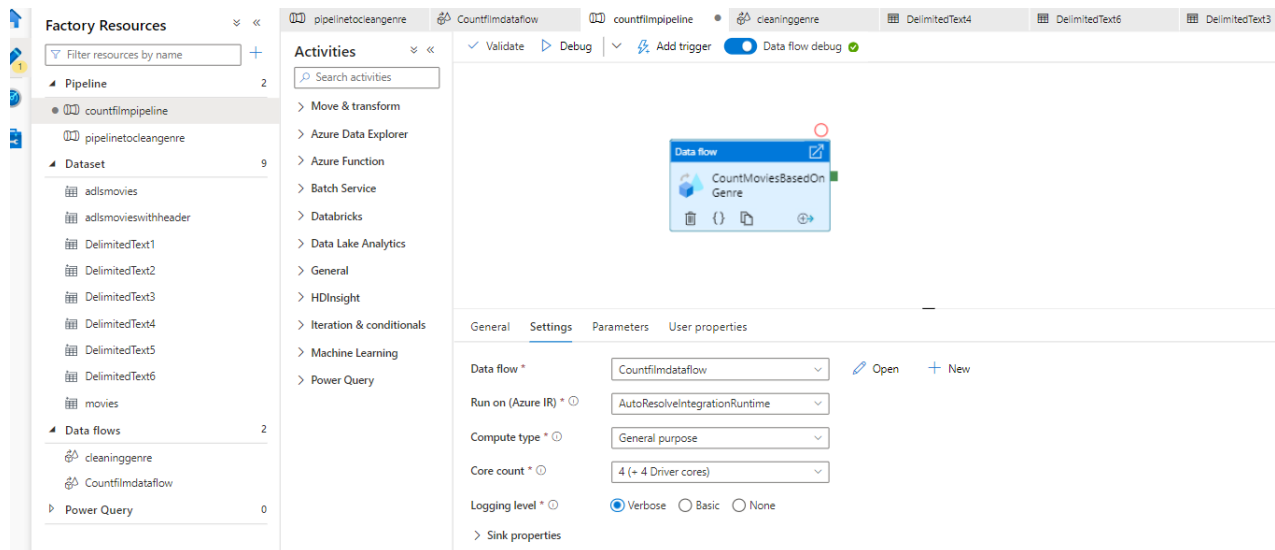
Film	Genre	LeadStudio	AudienceScore	Profitability	RottenTomatoes	WorldwideGross	Years
1	Romance	The Weinstein Company	70	1.747541667	64	\$41.94	2008
2	Comedy	The Weinstein Company	52	1.09	68	\$19.62	2010
3	Comedy	Independent	35	1.211818182	43	\$26.66	2010
4	Comedy	Disney	44	0	15	\$43.04	2010
5	Comedy	Fox	72	6.267647029	28	\$219.37	2008
6	Drama	20th Century Fox	72	3.081421053	60	\$117.09	2011
7	Animation	Disney	69	2.896919067	96	\$521.28	2008
8	Romance	Independent	67	11.0897415	89	\$32.18	2007
9	Romance	Independent	53	0.005	6	\$0.03	2011
10	Romance	Warner Bros.	54	4.184038462	17	\$217.57	2010
11	Romance	Independent	47	3.7241924	46	\$55.86	2007
12	Romance	Independent	68	6.383363636	26	\$702.17	2011
13	Romance	Summit	82	10.18002703	49	\$376.66	2008
14	Romance	Independent	68	5.402631579	14	\$205.20	2009
15	Romance	Summit	78	14.1964	27	\$703.62	2009
16	Romance	Summit	65	2.596205128	38	\$101.33	2009
17	Romance	Summit	74	7.8678	43	\$314.70	2009

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

## Data Flow : Countfilmdataflow



## Datapipeline: countfilmpipeline



## Output in ADLS

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease

Authentication method: Access key (Switch to Azure AD User Account)

Location: adffolder / solution / genreCount

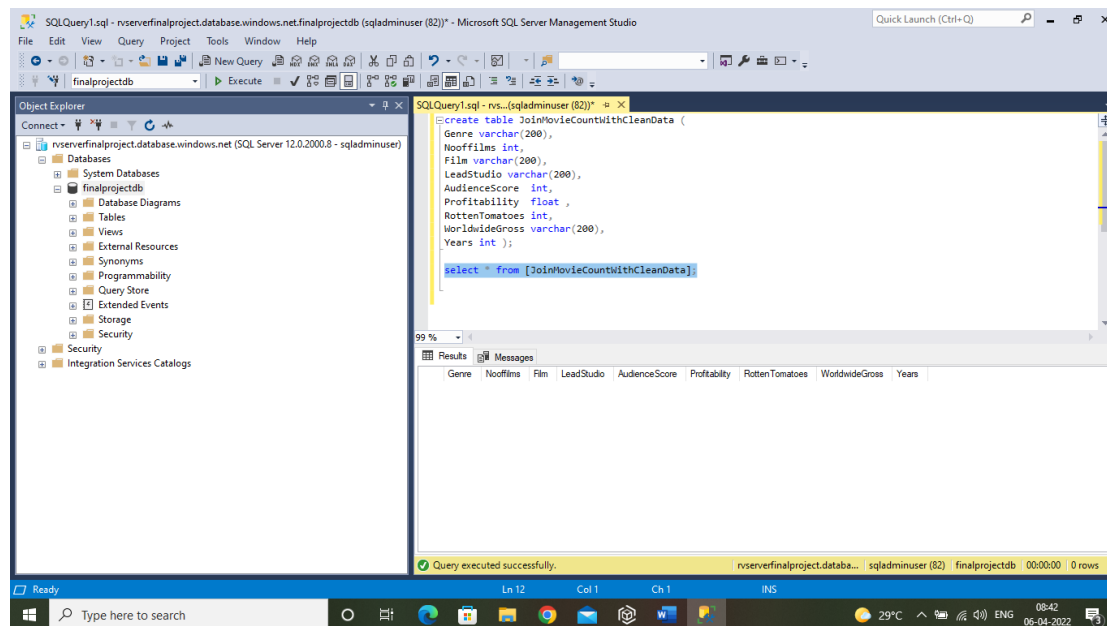
Search blobs by prefix (case-sensitive)

Name	Modified
[..]	
_committed_4372534847710359752	4/5/2022, 6:00:58 PM
_started_4372534847710359752	4/5/2022, 6:00:55 PM
_SUCCESS	4/5/2022, 6:00:58 PM
countoutput	4/5/2022, 8:41:34 PM

## Countoutput file

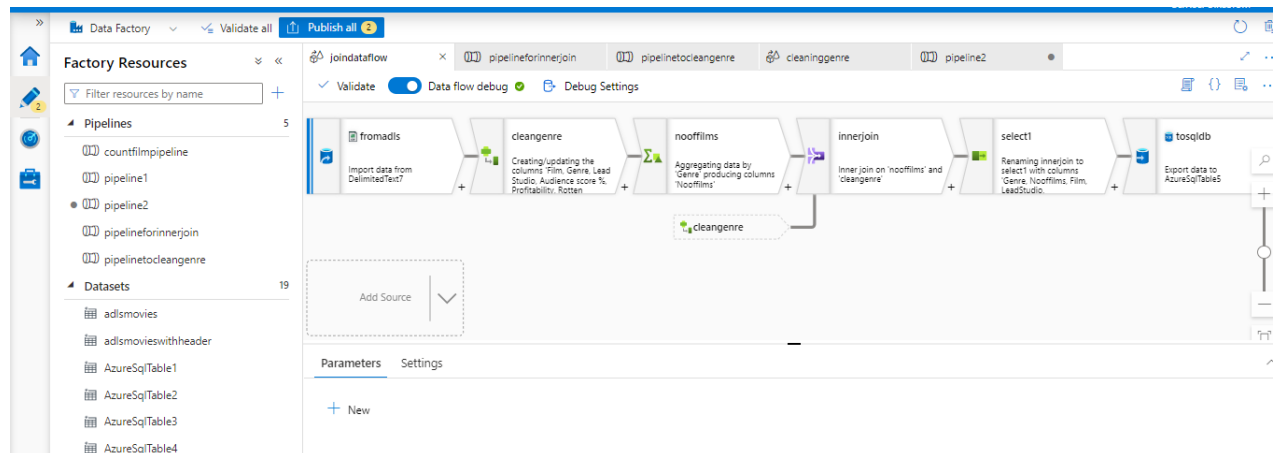
```
countoutput - Notepad
File Edit Format View Help
Genre,Film
Drama,13
Action,1
Fantasy,1
Comedy,43
Animation,4
Romance,15
```

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

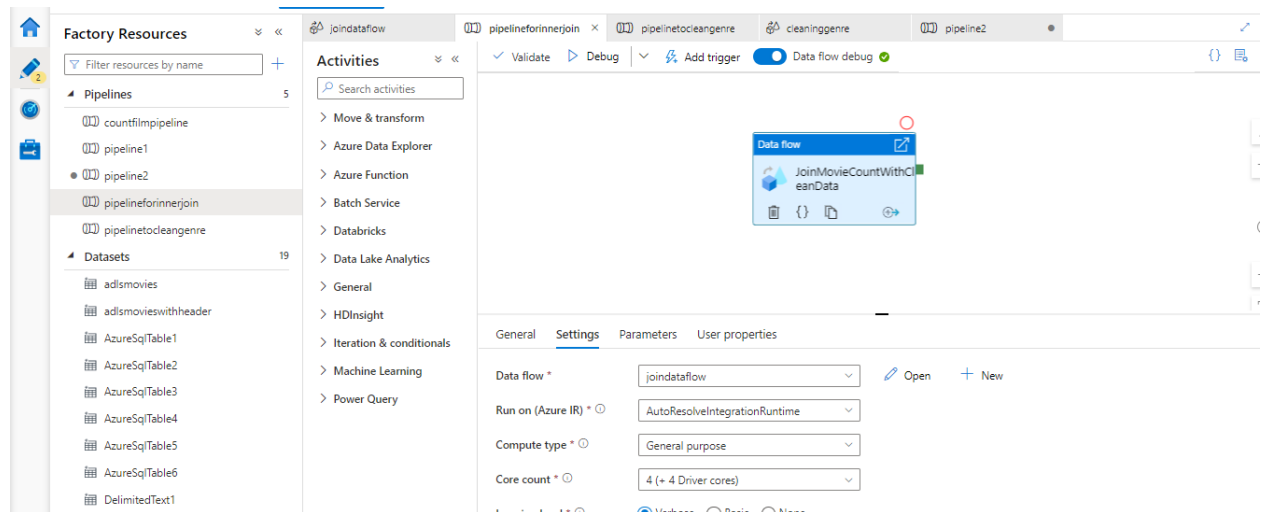


```
create table JoinMovieCountWithCleanData (
Genre varchar(200),
Nooffilms int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
Profitability float,
RottenTomatoes int,
WorldwideGross varchar(200),
Years int );
```

DataFlow : joindataflow



Pipeline : pipelineforinnerjoin



Output in SQL

Database:

Genre	Nooffilms	Film	LeadStudio	AudienceScore	Profitability	RottenTomatoes	WorldwideGross	Years
Romance	15	Zack and Mr Make a Porno	The Weinstein Company	70	1.747541667	64	\$41.94	2008
Comedy	43	Youth in Revolt	The Weinstein Company	52	1.09	68	\$19.52	2010
Comedy	43	You Will Meet a Tall Dark Stranger	Independent	35	1.211818182	43	\$26.66	2010
Comedy	43	When in Rome	Disney	44	0	15	\$43.04	2010
Comedy	43	What Happens in Vegas	Fox	72	6.267647029	28	\$219.37	2008
Drama	13	Water For Elephants	20th Century Fox	72	3.081421053	60	\$117.09	2011
Animation	4	WALL-E	Disney	89	2.896019067	96	\$521.28	2008
Romance	15	Watress	Independent	67	11.0897415	89	\$22.18	2007
Romance	15	Waiting For Forever	Independent	53	0.005	6	\$0.59	2011
Comedy	43	Valentine's Day	Warner Bros.	54	4.194038462	17	\$217.57	2010
Romance	15	Tyler Perry's Why Did I Get Married	Independent	47	3.7241924	46	\$55.86	2007
Romance	15	Twilight: Breaking Dawn	Independent	68	6.383363636	26	\$702.17	2011
Romance	15	Twilight	Summit	82	10.18002703	49	\$376.66	2008
Comedy	43	The Ugly Truth	Independent	68	5.402631579	14	\$205.30	2009
Drama	13	The Twilight Saga: New Moon	Summit	78	14.1964	27	\$759.82	2009
Drama	13	The Time Traveler's Wife	Paramount	65	2.598205128	38	\$101.33	2009
Comedy	43	The Proposal	Disney	74	7.8675	43	\$314.70	2009
Comedy	43	The Invention of Lying	Warner Bros.	47	1.751351351	56	\$32.40	2009
Comedy	43	The Heatseeker Kid	Paramount	41	2.129444167	30	\$127.77	2007

**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 toADLS Gen2.

## Datapipeline :

The screenshot shows the Microsoft Azure Data Factory console. On the left, the 'Factory Resources' pane lists pipelines and datasets. The 'Activities' pane shows various activity types. The main workspace displays a 'Copy data' activity configuration. The 'Source' tab is active, showing the 'Source dataset' as 'AzureSqlTable7'. The 'Use query' option is set to 'Table'. The 'Query timeout (minutes)' is set to 120. The 'Isolation level' is set to 'None'. The 'Partition option' is set to 'None'. A note at the bottom states: 'Please preview data to validate the partition settings are correct before you trigger a run or publish the pipeline.'

## Output in ADLS :

The screenshot shows the Azure Storage Explorer interface. The 'Overview' pane is active, displaying a table of blobs. The table has columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The table contains one row: 'dbo.movies.btt' with a modified date of '4/6/2022, 9:48:37 AM', an access tier of 'Hot (inferred)', a blob type of 'Block blob', a size of '6.13 KiB', and a lease state of 'Available'.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
dbo.movies.btt	4/6/2022, 9:48:37 AM	Hot (inferred)		Block blob	6.13 KiB	Available

## Output File:

```
dbo.movies - Notepad
File Edit Format View Help
Film,Genre,LeadStudio,AudienceScore,Profitability,RottenTomatoes,WorldwideGross,Years
"Zack and Miri Make a Porno","Romance","The Weinstein Company",70,1.7475416669999999,64,"$41.94 ",2008
"Youth in Revolt","Comedy","The Weinstein Company",52,1.0900000000000001,68,"$19.62 ",2010
"You Will Meet a Tall Dark Stranger","Comedy","Independent",35,1.211818182,43,"$26.66 ",2010
"When in Rome","Comedy","Disney",44,0,15,"$43.04 ",2010
"What Happens in Vegas","Comedy","Fox",72,6.2676470289999999,28,"$219.37 ",2008
"Water For Elephants","Drama","20th Century Fox",72,3.0814210530000001,60,"$117.09 ",2011
"WALL-E","Animation","Disney",89,2.8960190670000001,96,"$521.28 ",2008
"Waitress","Romance","Independent",67,11.089741500000001,89,"$22.18 ",2007
"Waiting For Forever","Romance","Independent",53,0.005000000000000001,6,"$0.03 ",2011
"Valentine's Day","Comedy","Warner Bros.",54,4.1840384620000002,17,"$217.57 ",2010
"Tyler Perry's Why Did I get Married","Romance","Independent",47,3.7241924000000002,46,"$55.86 ",2007
"Twilight: Breaking Dawn","Romance","Independent",68,6.3836363600000003,26,"$702.17 ",2011
"Twilight","Romance","Summit",82,10.18002703,49,"$376.66 ",2008
"The Ugly Truth","Comedy","Independent",68,5.4026315790000004,14,"$205.30 ",2009
"The Twilight Saga: New Moon","Drama","Summit",78,14.196400000000001,27,"$709.82 ",2009
"The Time Traveler's Wife","Drama","Paramount",65,2.598205128,38,"$101.33 ",2009
"The Proposal","Comedy","Disney",74,7.8674999999999997,43,"$314.70 ",2009
"The Invention of Lying","Comedy","Warner Bros.",47,1.7513513510000001,56,"$32.40 ",2009
"The Heartbreak Kid","Comedy","Paramount",41,2.1294441669999999,30,"$127.77 ",2007
"The Duchess","Drama","Paramount",68,3.2078502219999998,60,"$43.31 ",2008
"The Curious Case of Benjamin Button","Fantasy","Warner Bros.",81,1.7839437499999999,73,"$285.43 ",2008
"The Back-up Plan","Comedy","CBS",47,2.2025714289999998,20,"$77.09 ",2010
"Tangled","Animation","Disney",88,1.3656923080000001,89,"$355.01 ",2010
"Something Borrowed","Romance","Independent",48,1.7195142859999999,15,"$60.18 ",2011
"She's Out of My League","Comedy","Paramount",60,2.4405000000000001,57,"$48.81 ",2010
"Sex and the City Two","Comedy","Warner Bros.",49,2.8835000000000002,15,"$288.35 ",2010
"Sex and the City 2","Comedy","Warner Bros.",49,2.8835000000000002,15,"$288.35 ",2010
"Sex and the City","Comedy","Warner Bros.",81,7.2217957909999999,49,"$415.25 ",2008
"Remember Me","Drama","Summit",70,3.49125,28,"$55.86 ",2010
"Rachel Getting Married","Drama","Independent",61,1.3841666669999999,85,"$16.61 ",2008
"Penelope","Comedy","Summit",74,1.3827997329999999,52,"$20.74 ",2008
"P.S. I Love You","Romance","Independent",82,5.1031168329999996,21,"$153.09 ",2007
"Over Her Dead Body","Comedy","New Line",47,2.0710000000000002,15,"$20.71 ",2008
"Our Family Wedding","Comedy","Independent",49,0,14,"$21.37 ",2010
"One Day","Romance","Independent",54,3.6827333329999998,37,"$55.24 ",2011
```

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.

```
account_name = "rvadlsfinalproject"
container_name = "adlstoadls"
input_relative_path = "input"
output_relative_path = "output"
adls_path = 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name, account_name, input_relative_path)
adls_output_path= 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name,
    account_name, output_relative_path)
```

```
spark.conf.set("fs.azure.account.auth.type.%s.dfs.core.windows.net" %account_name, "SharedKey")
spark.conf.set("fs.azure.account.key.%s.dfs.core.windows.net" %account_name
    , "HwZ0JciEFkqRHSNQGyECXjD+94YMR5SJMwjDJ5E/MlrgBebJa+jx7q3PWxqmKNN/H6bGjAvziNrB+
    AS1x1PzSA==")
```

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType
txnSchema = StructType([
    StructField("Film", StringType(), True),
    StructField("Genre", StringType(), True),
    StructField("LeadStudio", StringType(), True),
    StructField("AudienceScore", IntegerType(), True),
    StructField("Profitability", DoubleType(), True),
    StructField("RottenTomatoes", IntegerType(), True),
```

```
StructField("WorldwideGross",StringType(),True),
StructField("Years",IntegerType(),True),
```

```
)
```

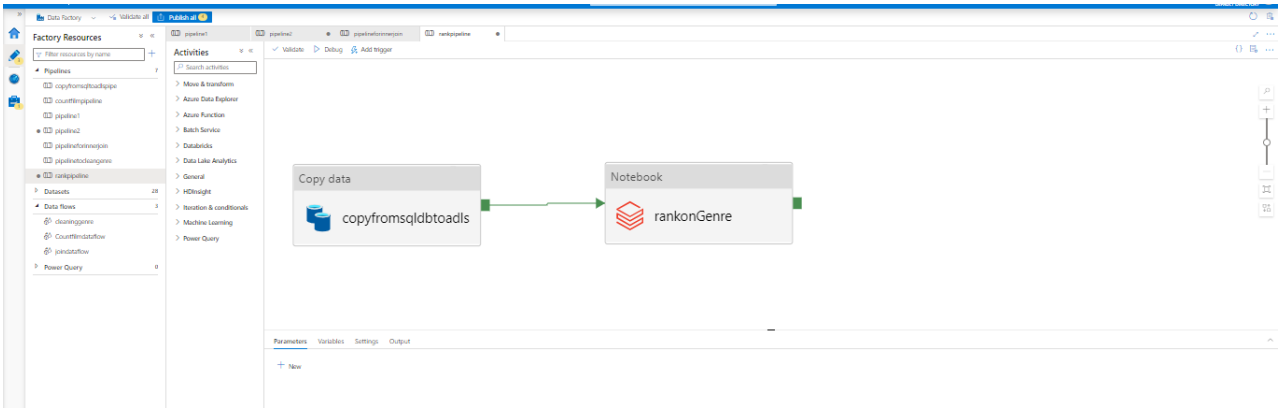
```
dfmovies = spark.read.option('header', 'true').option('delimiter', ',').schema(txnSchema).csv(adls_path)
```

```
dfmovies.registerTempTable("movies")
```

```
resultmovies = spark.sql("SELECT *, RANK () OVER (ORDER BY Genre) AS Rank_no FROM movies ")
```

```
resultmovies.write.option('header', 'true').option('delimiter', ',').csv(adls_output_path)
```

### Pipeline:



### Output in ADLS :

The screenshot shows the Microsoft Azure portal interface for the 'adlstoaddls' container. The 'Overview' tab is selected, displaying a table of blobs. The table includes columns for Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The following table represents the data shown in the screenshot:

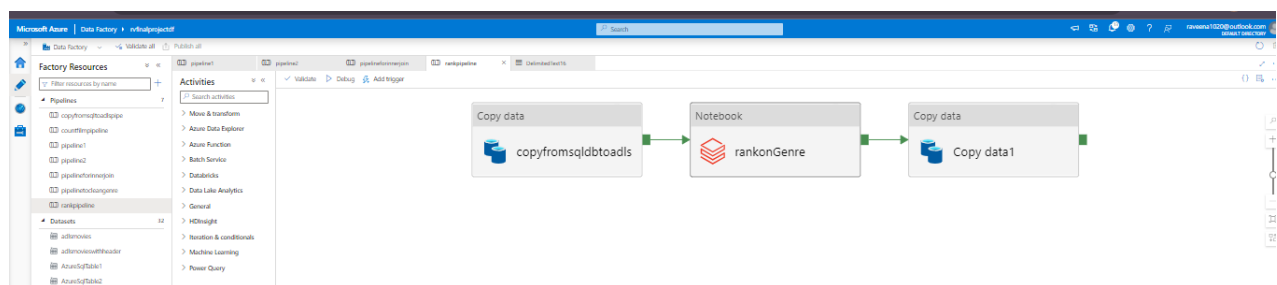
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[.]						
_committed_231577776784933247	4/6/2022, 2:10:48 PM	Hot (Inferred)		Block blob	111 B	Available
_started_231577776784933247	4/6/2022, 2:10:48 PM	Hot (Inferred)		Block blob	0 B	Available
_SUCCESS	4/6/2022, 2:10:48 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-tid-231577776784933247-ab084ef5-bac3-44dd-afef-6409e036d5eb-2b-1-c000.csv	4/6/2022, 2:10:48 PM	Hot (Inferred)		Block blob	5.56 KiB	Available

Film	Genre	LeadStudio	Audience	Profitability	RottenTomatoes	WorldwideGross	Years	Rank_no
1	Killers	Action	Lionsgate	45	1.245333	11 \$93.40	2010	1
2	WALL-E	Animation	Disney	89	2.896019	96 \$521.28	2008	2
3	Tangled	Animation	Disney	88	1.365692	89 \$355.01	2010	2
4	Gnomeo & Juliet	Animation	Disney	52	5.387972	56 \$193.97	2011	2
5	Gnomeo & Juliet	Animation	Disney	52	5.387972	56 \$193.97	2011	2
6	You Will Meet a Tall Dark Stranger	Comedy	The Weinstein Company	35	1.211818	43 \$26.66	2010	6
7	You Will Meet a Tall Dark Stranger	Comedy	Independent	35	1.211818	43 \$26.66	2010	6
8	When in Rome	Comedy	Disney	44	0	15 \$43.04	2010	6
9	What Happens in Vegas	Comedy	Fox	72	6.267647	28 \$219.37	2008	6
10	Valentine's Day	Comedy	Warner Bros.	54	4.184038	17 \$217.57	2010	6
11	The Ugly Truth	Comedy	Independent	68	5.402632	14 \$205.30	2009	6
12	The Proposal	Comedy	Disney	74	7.8675	43 \$314.70	2009	6
13	The Invention of Solitude	Comedy	Warner Bros.	47	1.751351	56 \$32.40	2009	6
14	The Heart of Christmas	Comedy	Paramount	41	2.129444	30 \$127.77	2007	6
15	The Backlist	Comedy	CBS	47	2.202571	20 \$77.09	2010	6
16	She's Out of Control	Comedy	Paramount	60	2.4405	57 \$48.81	2010	6
17	Sex and the City	Comedy	Warner Bros.	49	2.8835	15 \$288.35	2010	6
18	Sex and the City	Comedy	Warner Bros.	49	2.8835	15 \$288.35	2010	6
19	Sex and the City	Comedy	Warner Bros.	81	7.221796	49 \$415.25	2008	6
20	Penelope	Comedy	Summit	74	1.3828	52 \$20.74	2008	6

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

```
create table movies(
Genre varchar(200),
FilmCount int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
Profitability float,
RottenTomatoes int,
WorldwideGross varchar(200),
Years int,
Rank_no int)
```

Pipeline : rankpipeline



Output in Synapse:

Genre	FilmCount	Film	LeadStudio	Audiencescore	Profitability	RottenTomatoes	WorldwideGross	Year	Rank_no
Animation	4	WALL-E	Disney	89	2.896019067	96	\$521.28	2008	2
Animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56	\$193.97	2011	2
Animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56	\$193.97	2011	2
Action	1	Killers	Lionsgate	45	1.245333333	11	\$93.40	2010	1
Animation	4	Tangled	Disney	88	1.365692308	89	\$355.01	2010	2
Comedy	43	You Will Meet a Tall Dark Stranger	Independent	35	1.211818182	43	\$26.66	2010	6
Comedy	43	What Happens in Vegas	Fox	72	6.267647029	28	\$219.37	2008	6
Comedy	43	The Ugly Truth	Independent	68	5.402631579	14	\$205.30	2009	6
Comedy	43	The Proposal	Disney	74	7.8675	43	\$314.70	2009	6
Comedy	43	When in Rome	Disney	44	0	15	\$43.04	2010	6
Comedy	43	Valentine's Day	Warner Bros.	54	4.184038462	17	\$217.57	2010	6

Query executed successfully. synapsefinalproject.sql.azur... sqladminuser (149) synapsesqlpool 00:00:00 77 rows