AZURE PROJECT

RAVEENA PRABAKARAN

TIGER ANALYTICS

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:

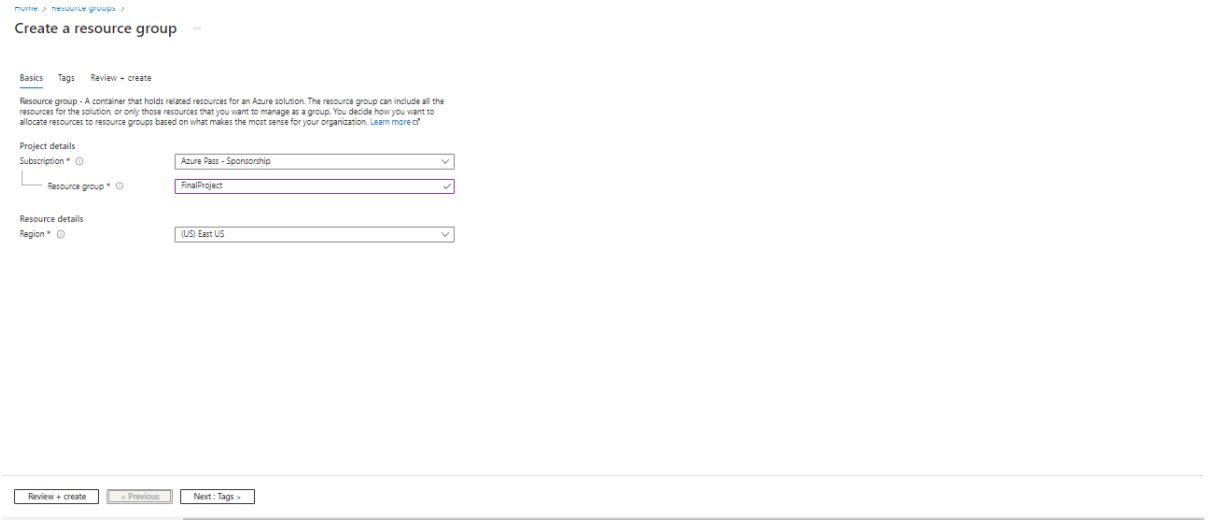
As a Data Engineer, you've been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will be collecting the data from SQL Database using ADF and the transformed data will be the source for databricks to run complex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

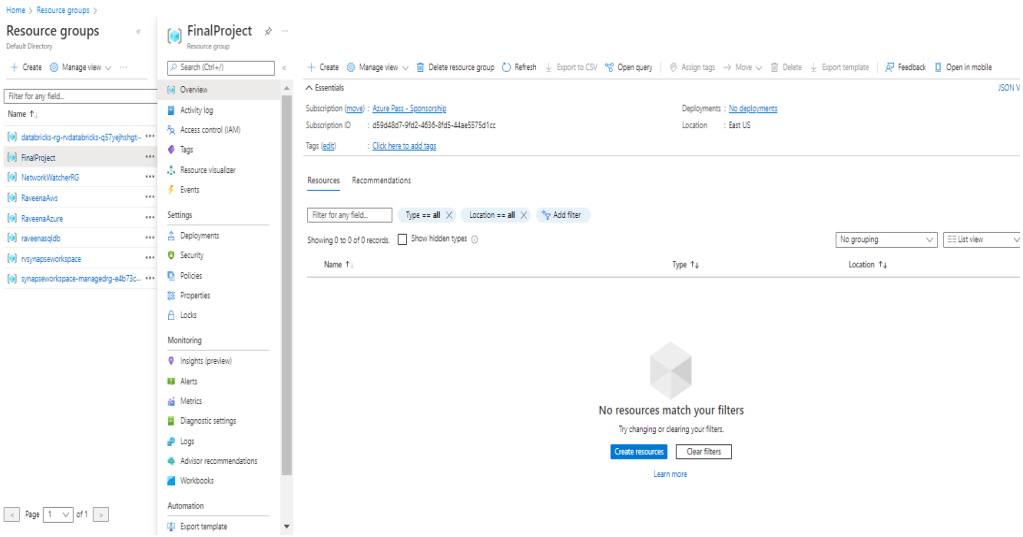
Steps for building ETL pipeline :

In this project, perform the following steps:

- Create a Resource Group.

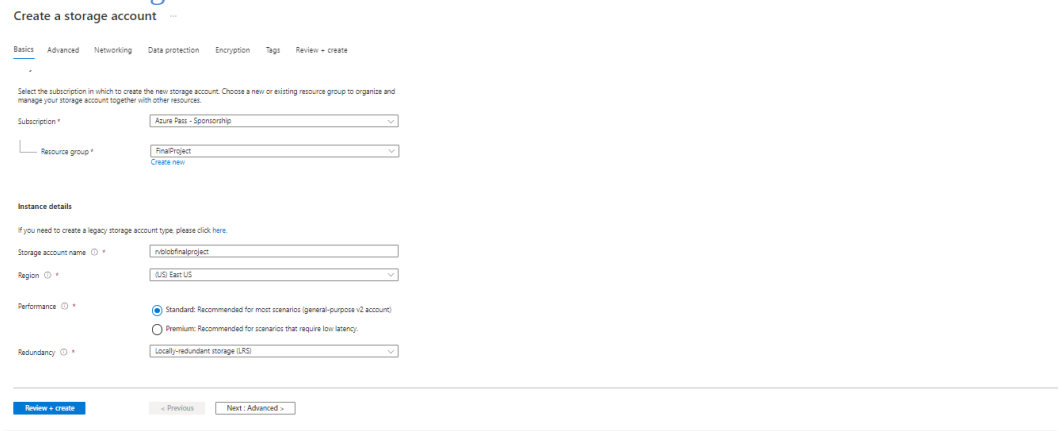


Resource Group → FinalProject

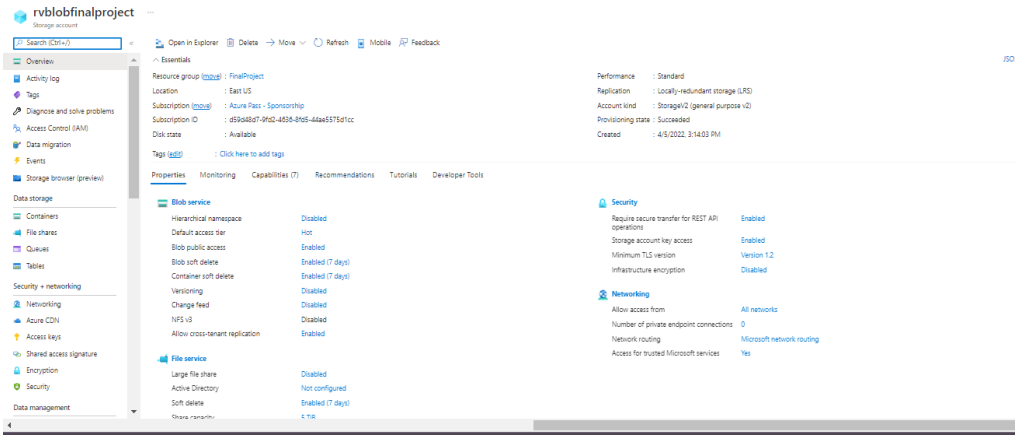


- ## Create a Storage account.

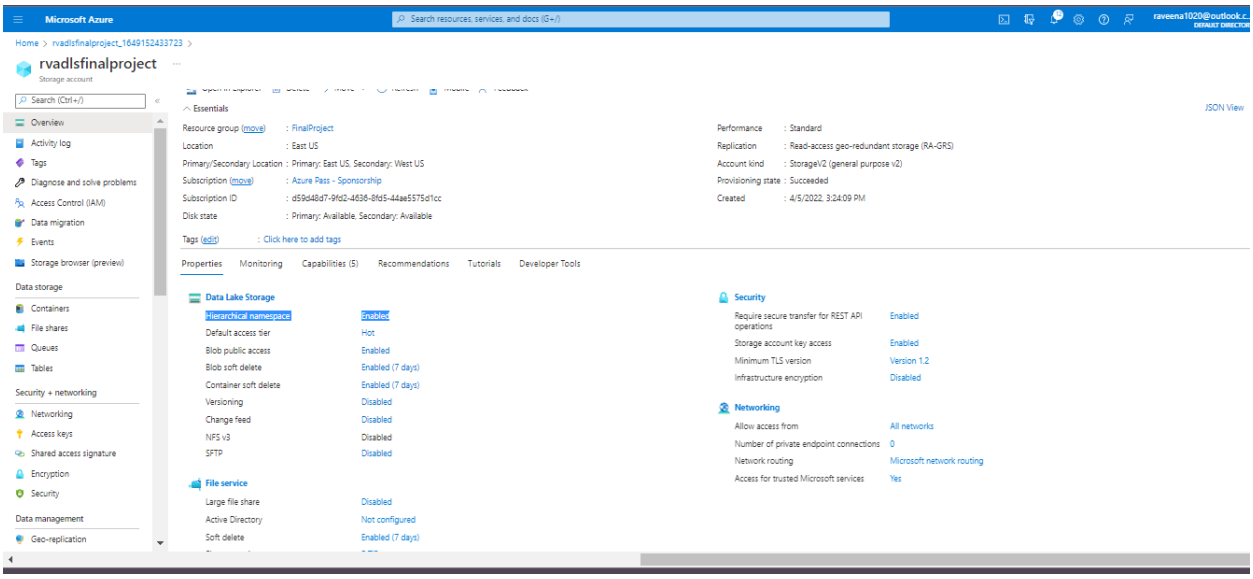
a) Blob Storage



Blob storage → rvblobfinalproject

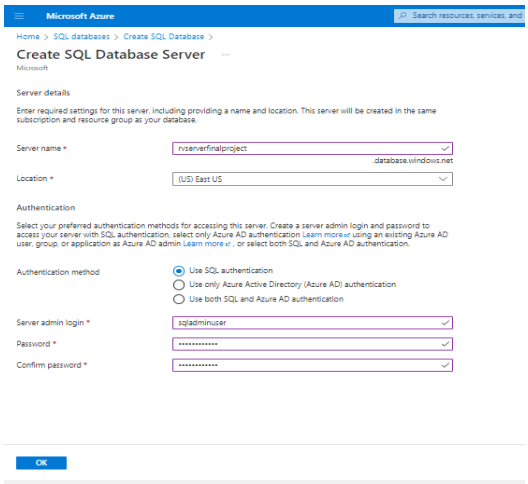


b) Azure Data Lake Storage → rvadlsfinalproject

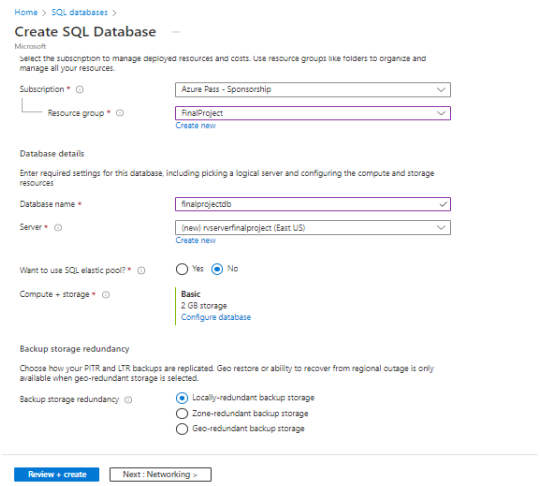


- Create an Azure SQL Database.

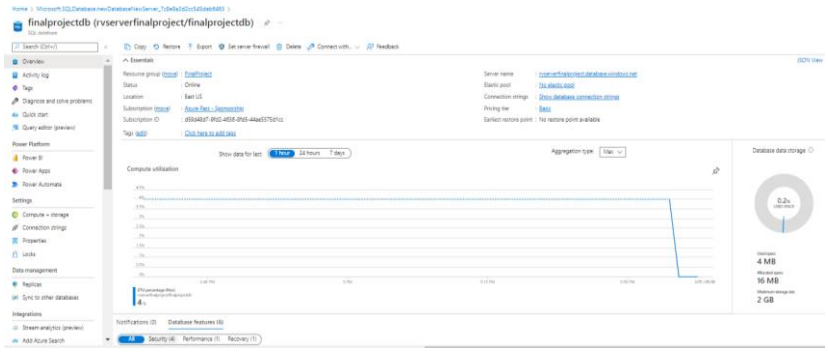
Server



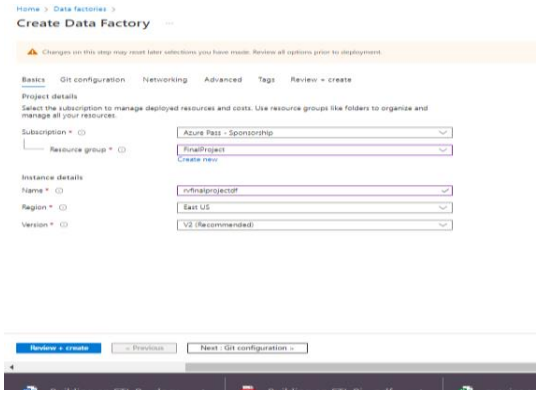
Sql Database



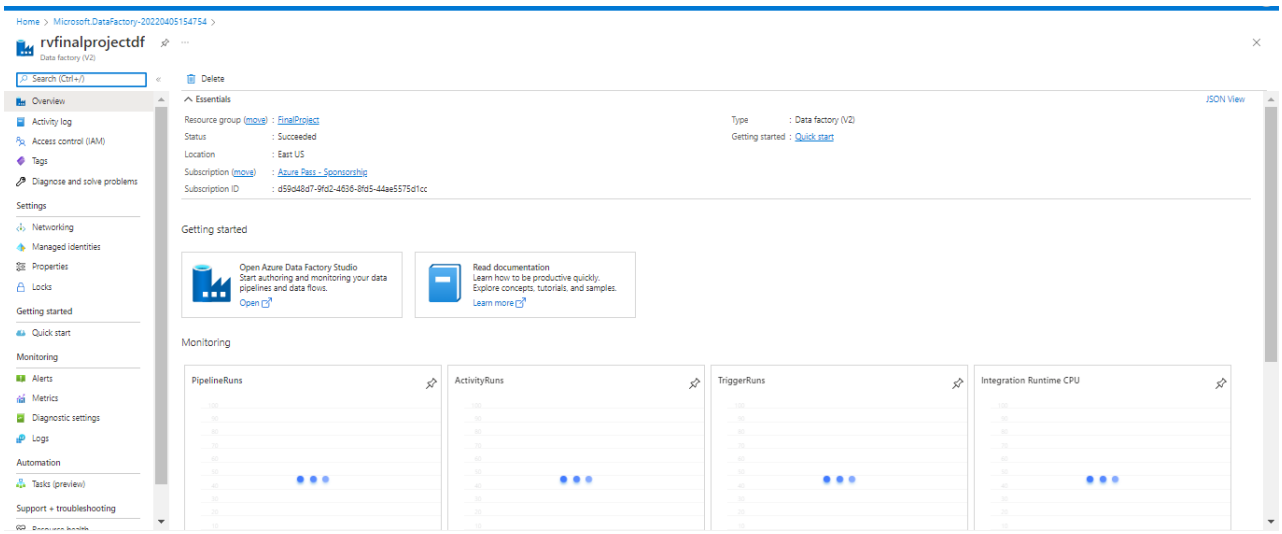
Sql Database --> finalprojectdb



- Create a data factory.

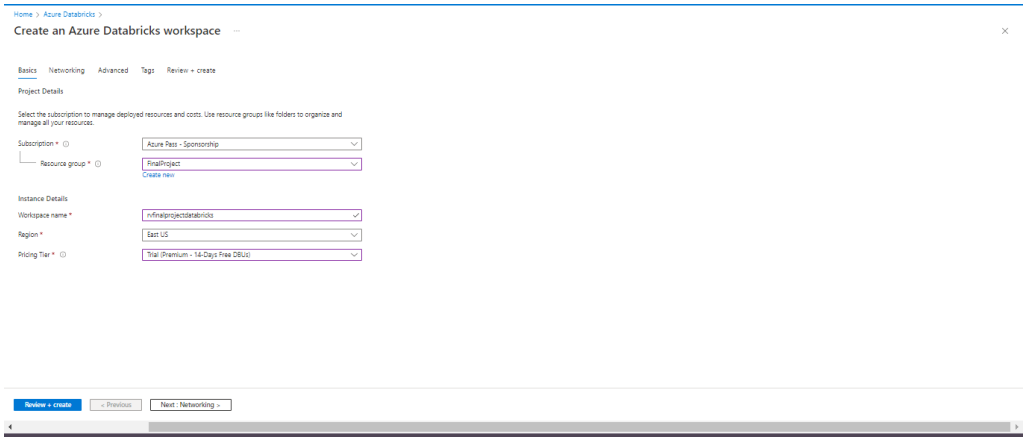


Data Factory → rvfinalprojectdf

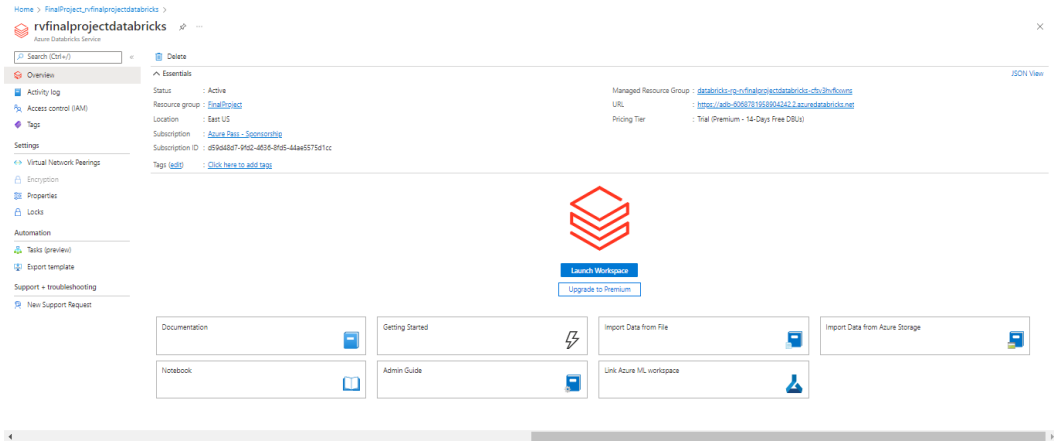


- Configure Databricks cluster

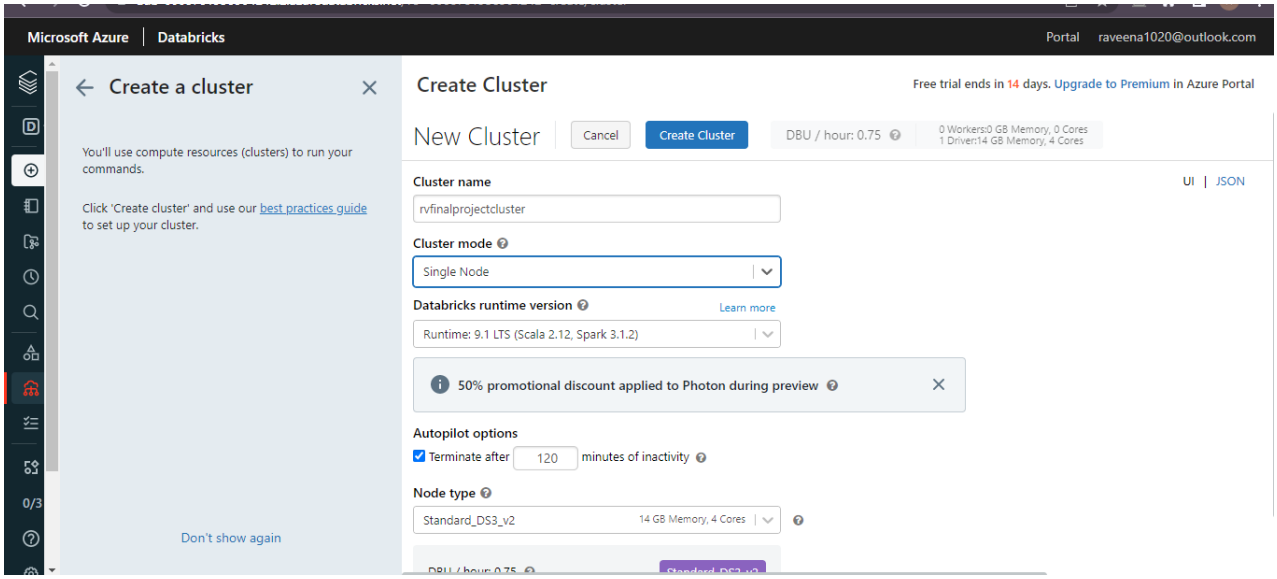
Creating Azure Databricks

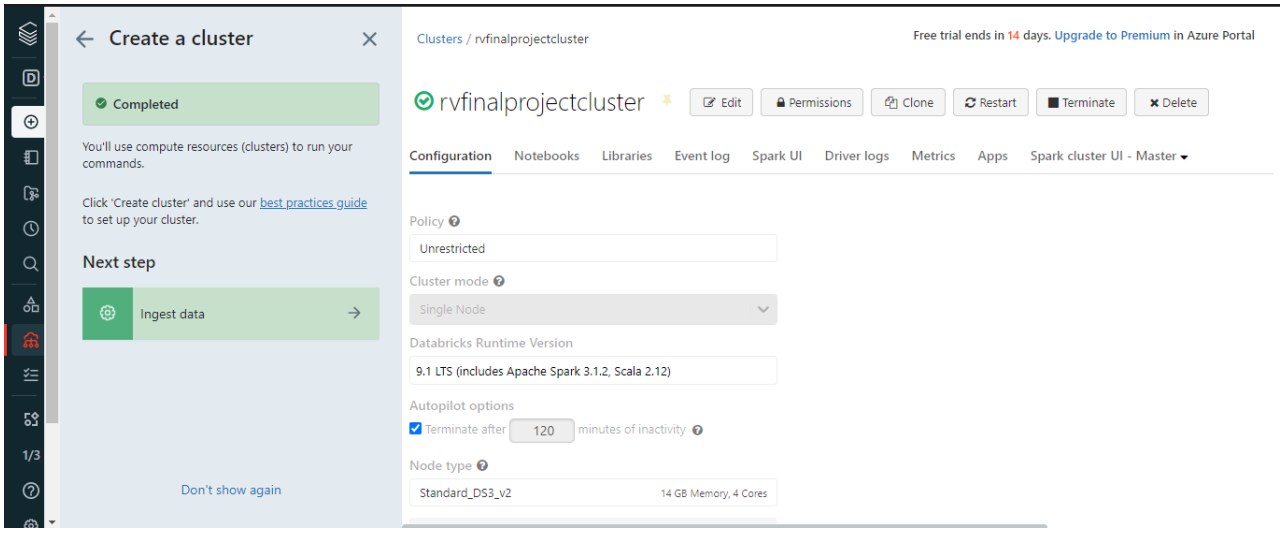


Databricks → rvfinalprojectdatabricks

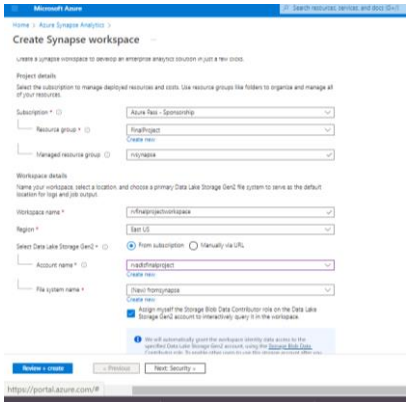


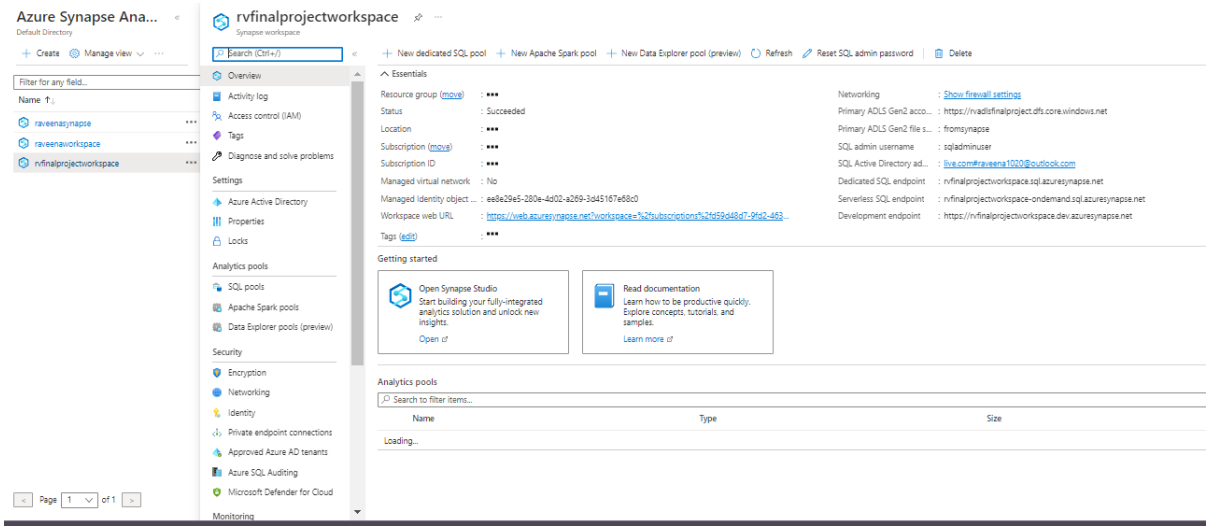
Creating Cluster → rvfinalprojectcluster



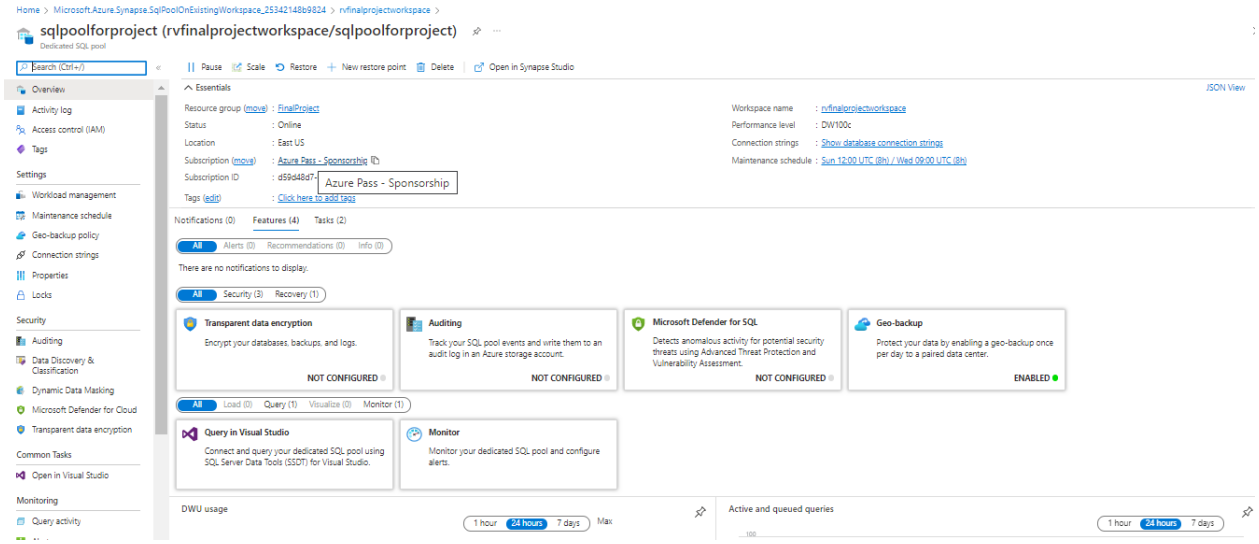


- Create Synapse analytics Data Warehouse.





SQLPOOL



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

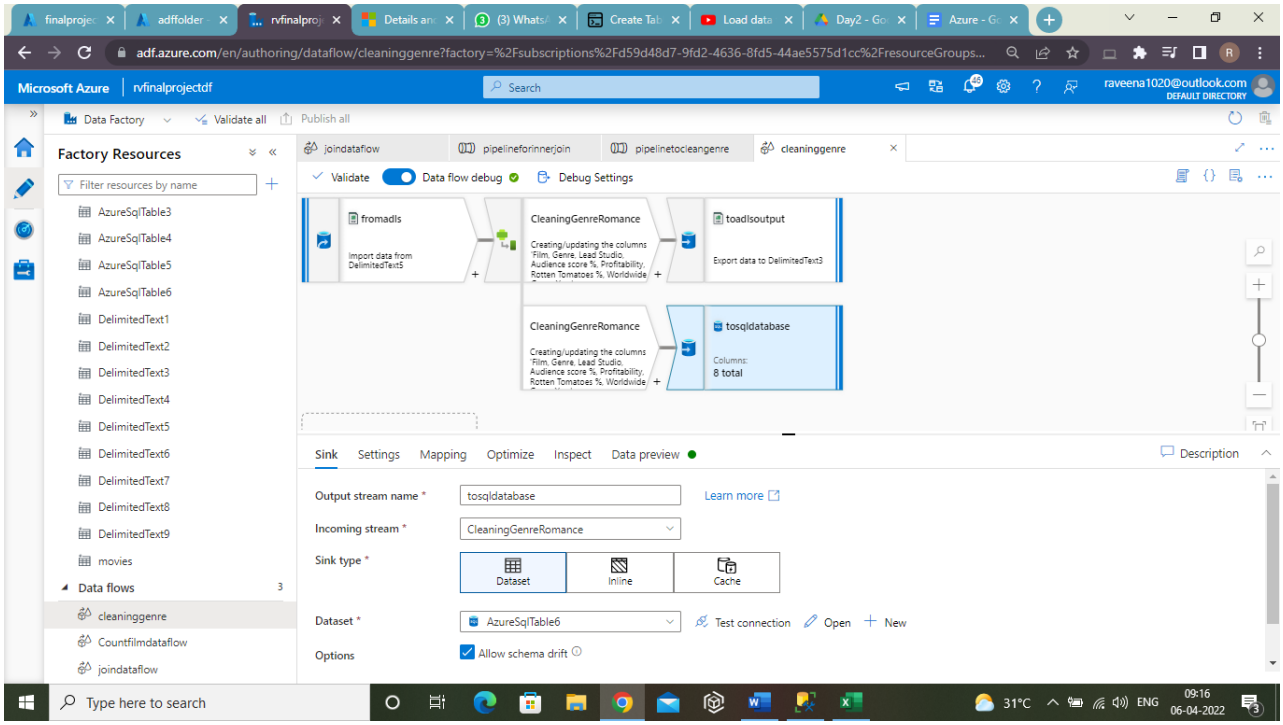
- Use Databricks notebook for mounting ADLS Gen 2 storage, transforming the data (clean, join, filter, aggregate, pivot) and persist result to ADLS.
- Schedule and Monitor the pipeline and activity runs.

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

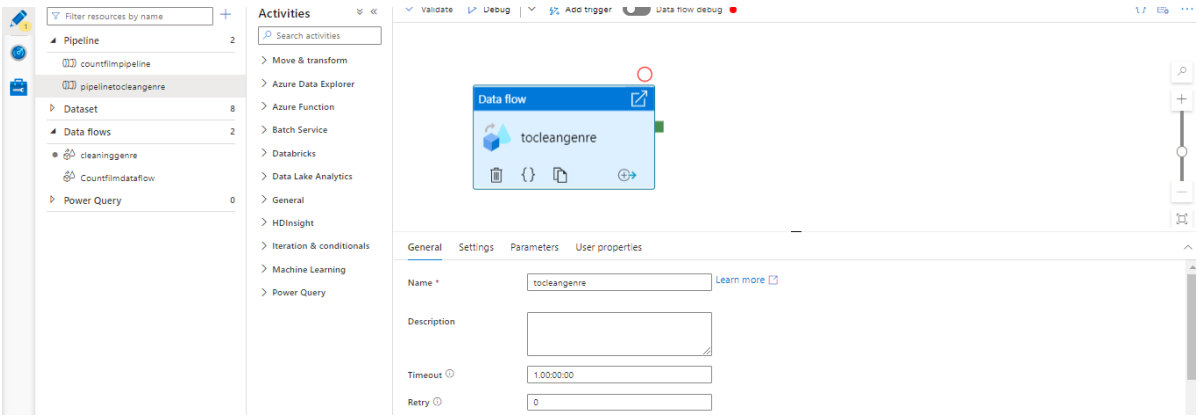
Task 1: Create a dataflow with the following requirement:

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.)

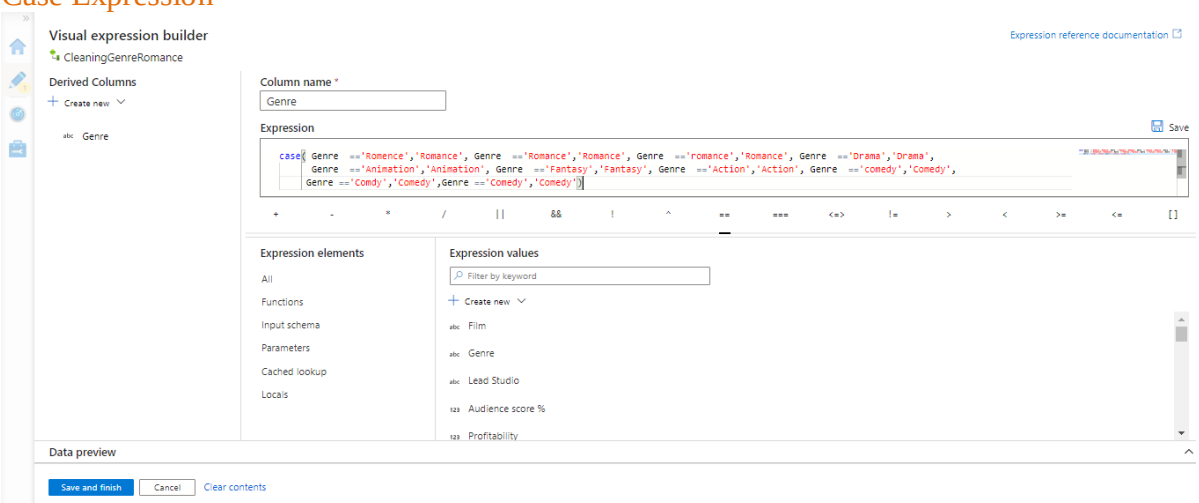
Dataflow cleaninggenre



Datapipeline pipelinetocleangenre

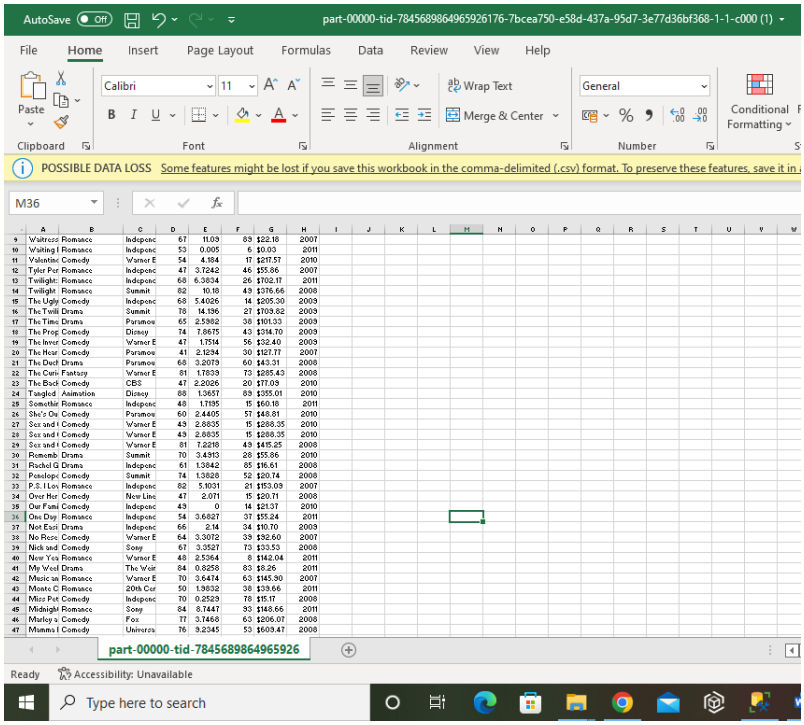


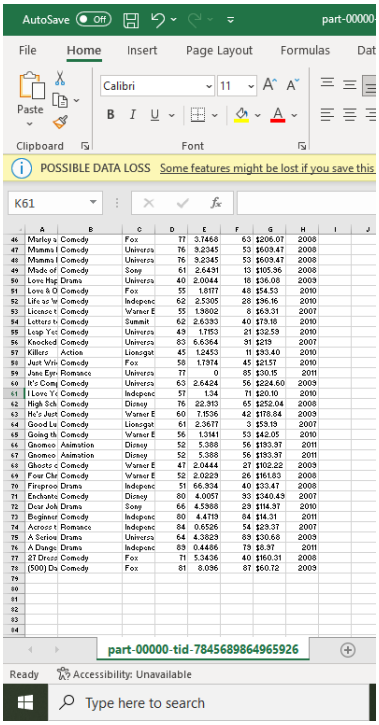
Case Expression



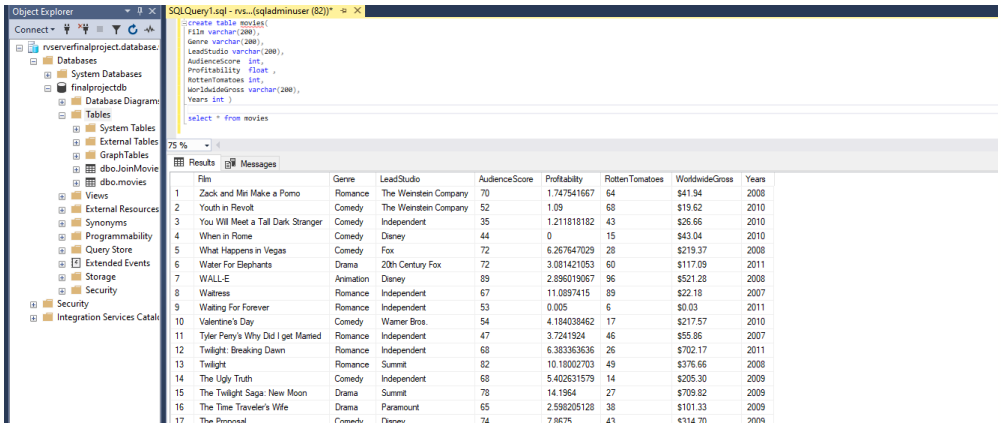

```
case( Genre  ='Romence','Romance', Genre  ='Romance','Romance', Genre  ='romance','Romance', Genre  ='Drama','Drama',Genre  ='Animation','Animation', Genre  ='Fantasy','Fantasy', Genre  ='Action','Action', Genre  ='comedy','Comedy',Genre ='Comdy','Comedy',Genre ='Comedy','Comedy')
```

Output in ADLS:



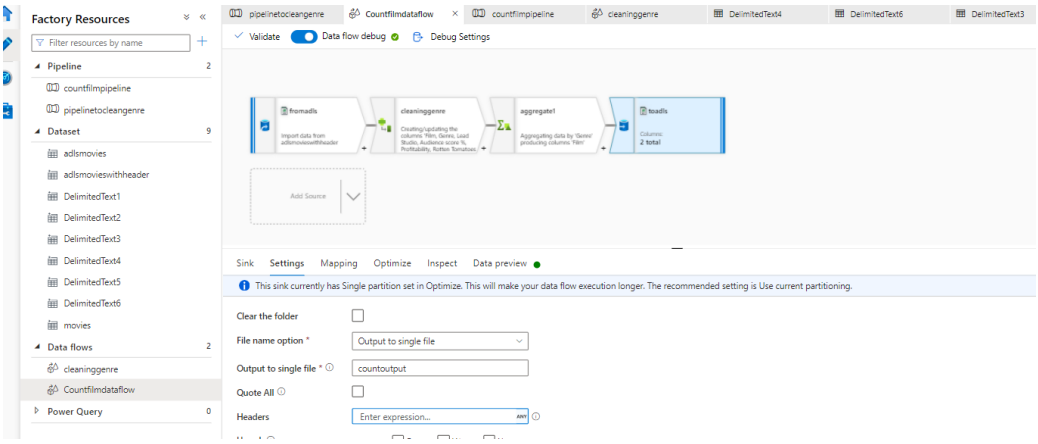


Output in Sql Database :

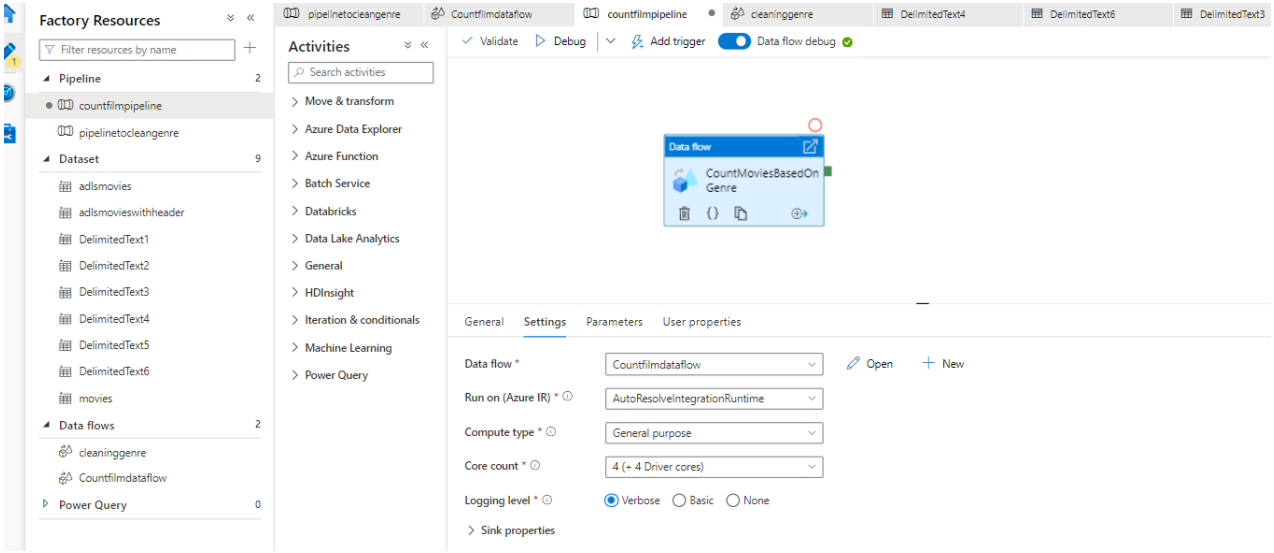


2. Create a data stream named CountMoviesBasedOnGenre that can calculate number of films foreach genre and store it as a separate dataset in ADLS under folder name “solution/genreCount”

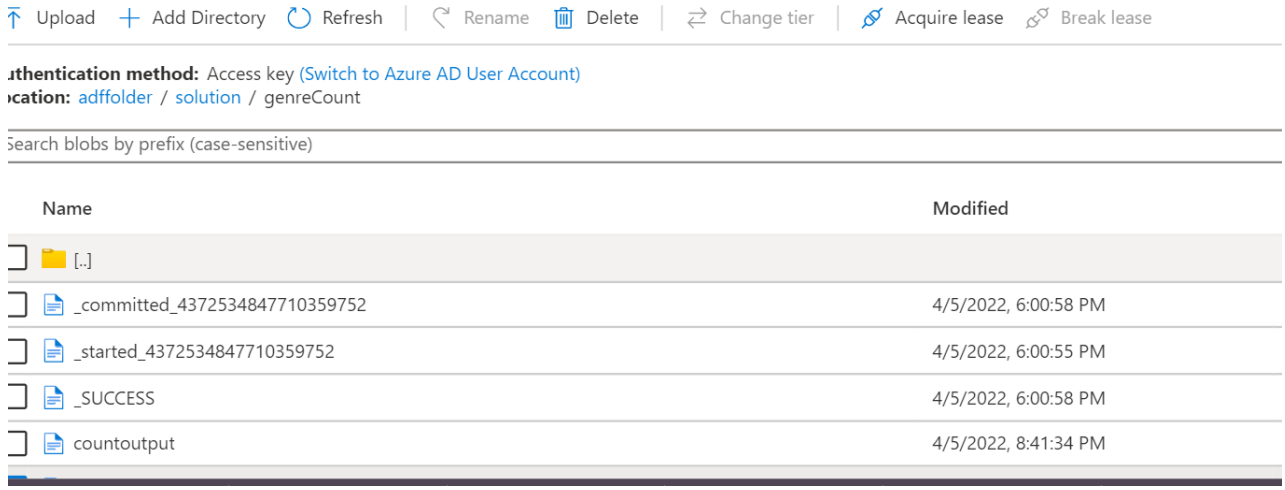
Data Flow : Countfilmdataflow



Datapipeline: countfilmpipeline



Output in ADLS



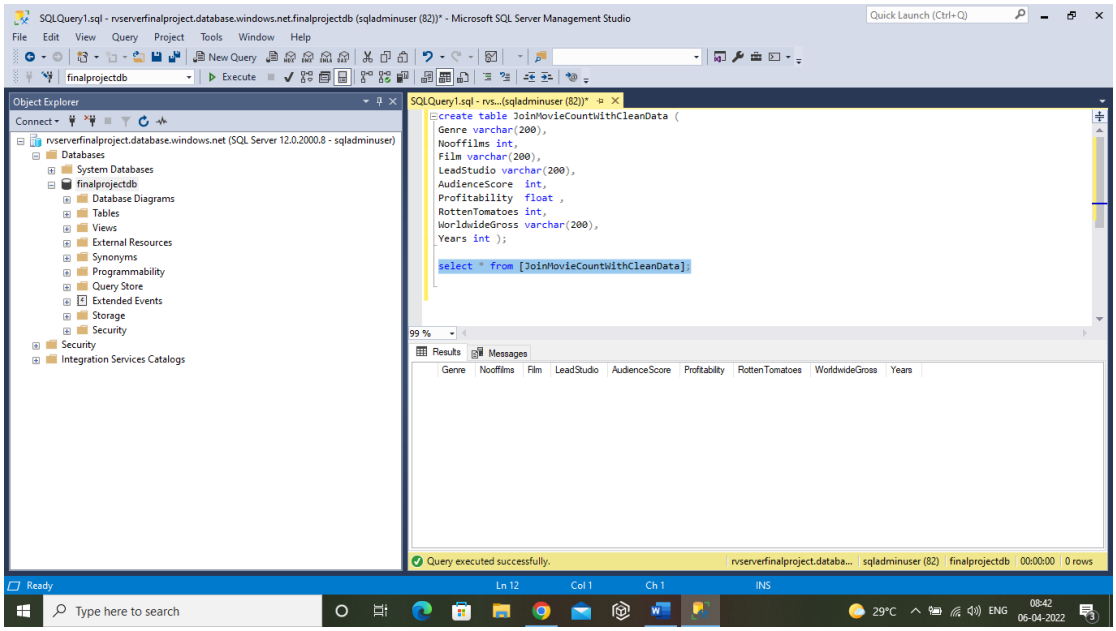
Countoutput file

countoutput - Notepad

File Edit Format View Help

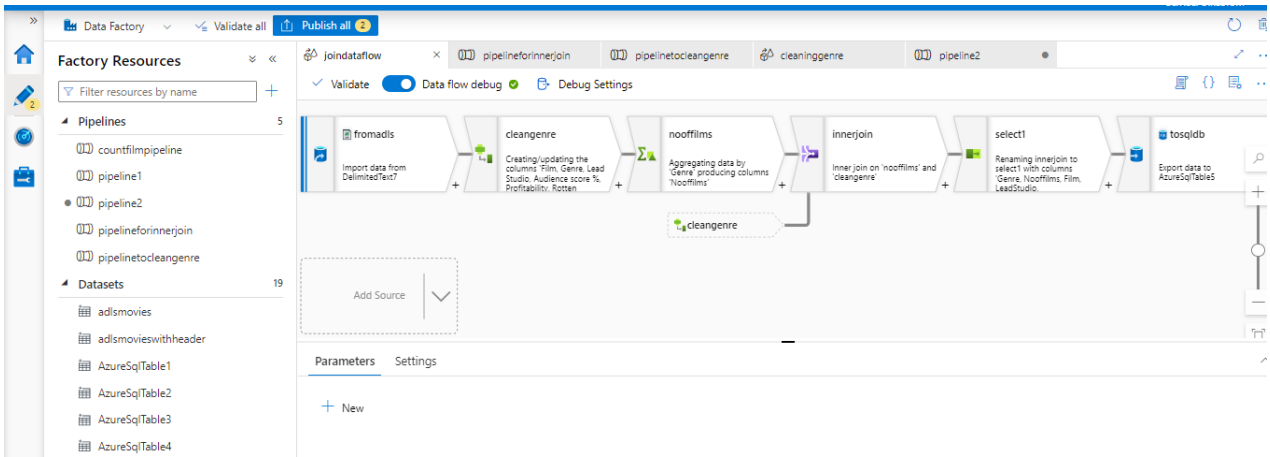
```
Genre,Film
Drama,13
Action,1
Fantasy,1
Comedy,43
Animation,4
Romance,15
```

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

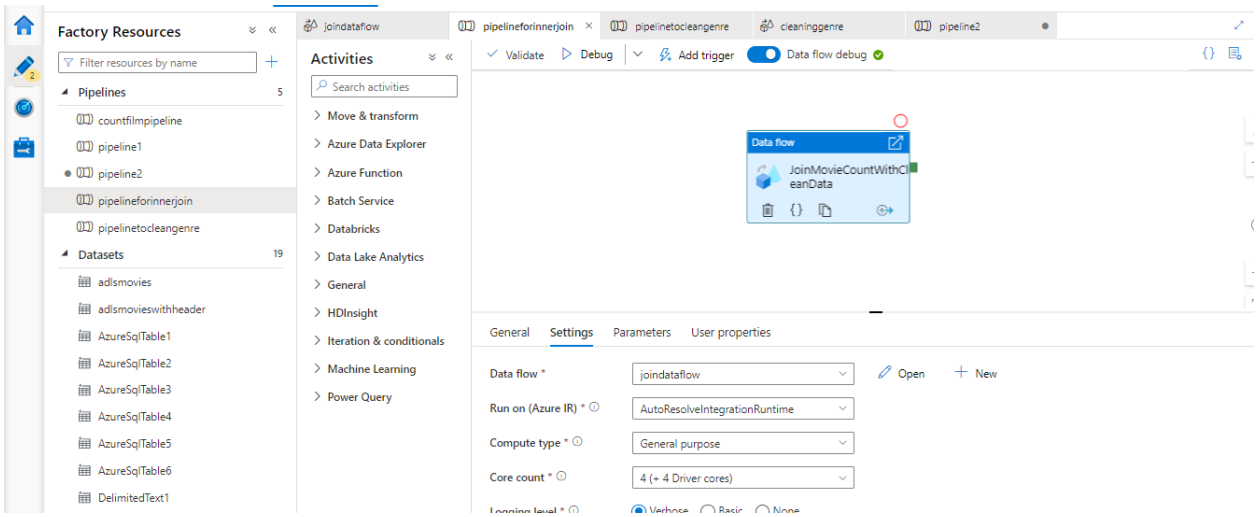



```
create table JoinMovieCountWithCleanData (
Genre varchar(200),
Nooffilms int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
Profitability  float ,
RottenTomatoes int,
WorldwideGross varchar(200),
Years int );
```

DataFlow : joindataflow

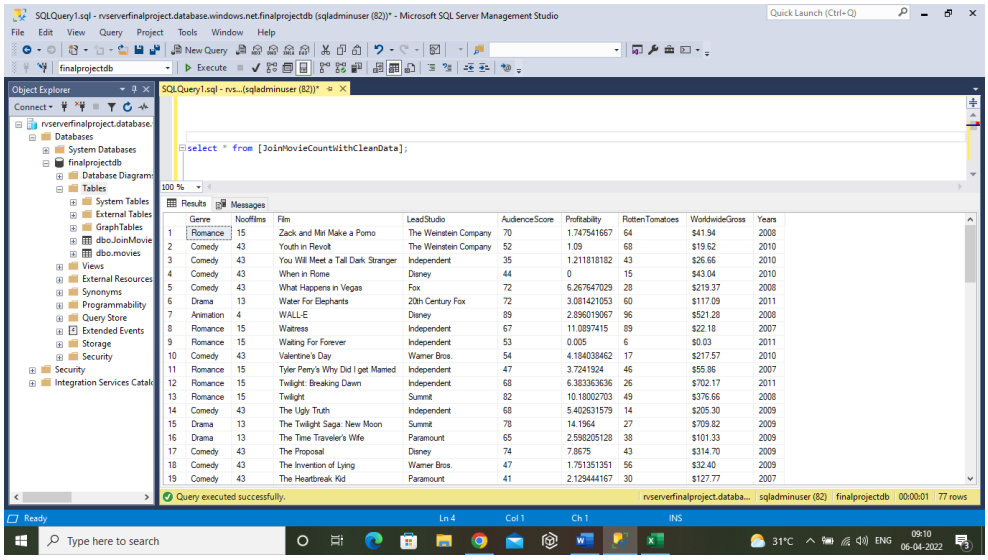


Pipeline : pipelineforinnerjoin



Outptut in SQL

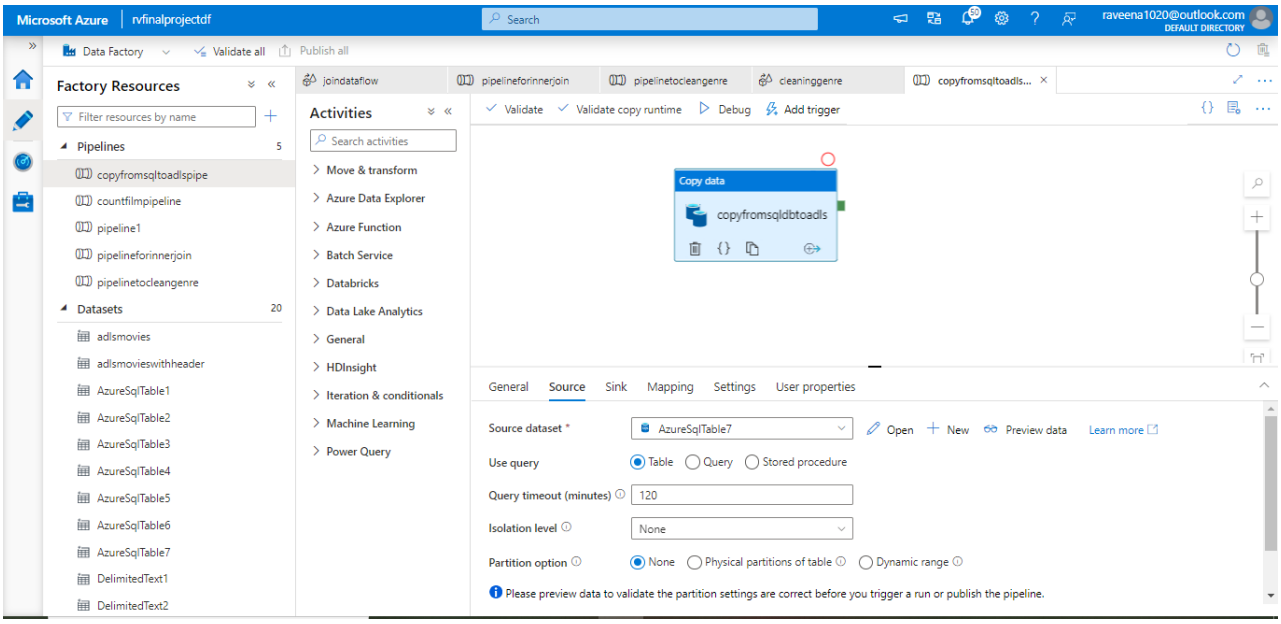
Database:



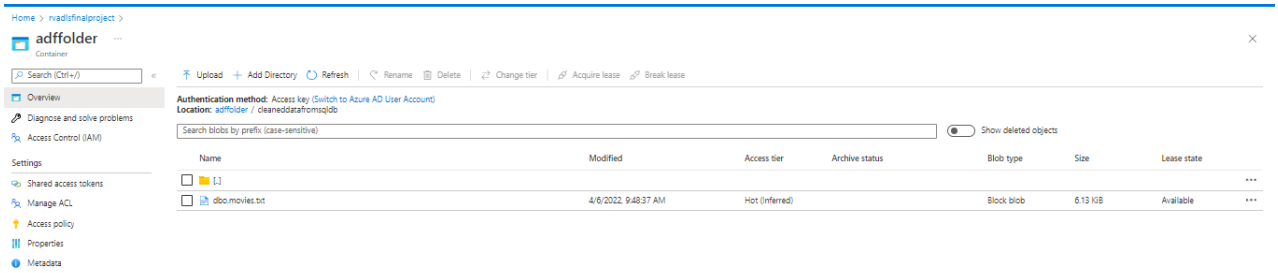
Task 2: Create the following activity pipeline

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

Datapipeline :



Output in ADLS :



Output File:

dbo.movies - Notepad

File Edit Format View Help

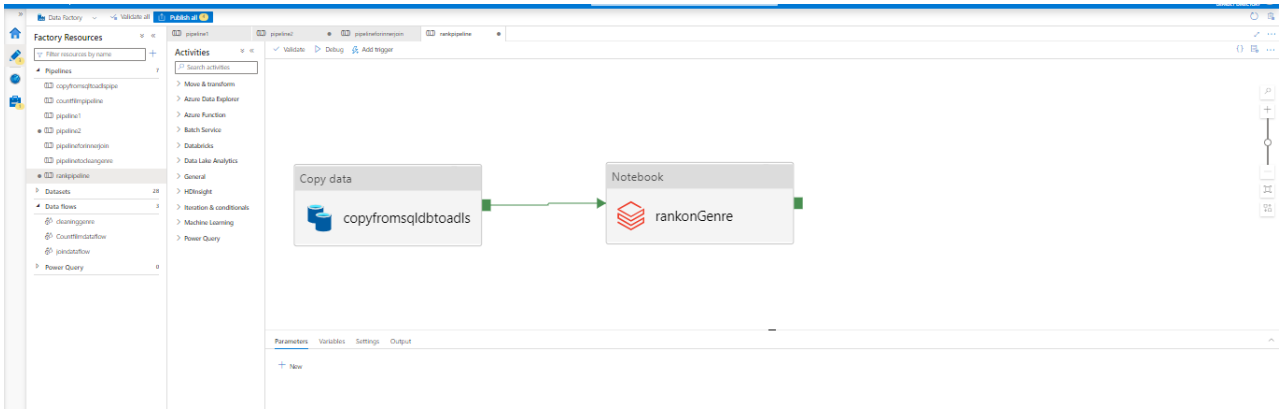
```
Film,Genre,LeadStudio,AudienceScore,Profitability,RottenTomatoes,WorldwideGross,Years
"Zack and Miri Make a Porno","Romance","The Weinstein Company",70,1.7475416669999999,64,"$41.94 ",2008
"Youth in Revolt","Comedy","The Weinstein Company",52,1.0900000000000001,68,"$19.62 ",2010
"You Will Meet a 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",53,0.0050000000000000001,6,"$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",68,6.3833636360000003,26,"$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 performrank 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.

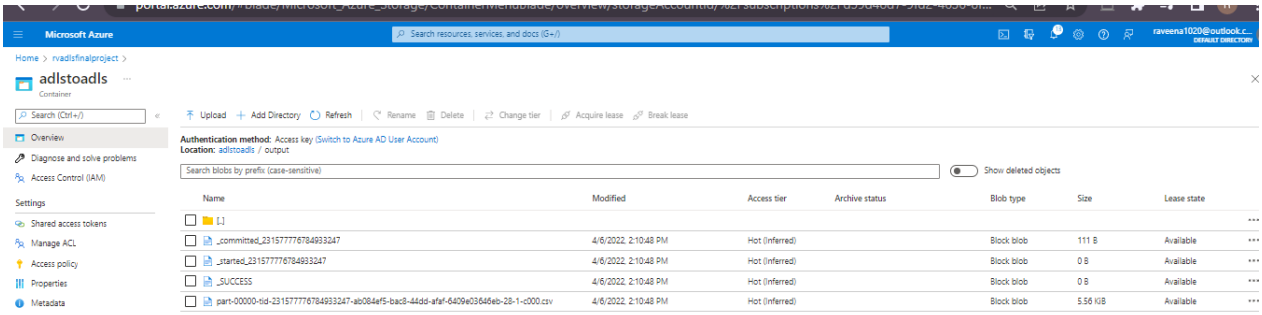
```
account_name = "rvadlsfinalproject"
container_name = "adlstoadls"
input_relative_path = "input"
output_relative_path = "output"
adls_path = 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name, account_name,input_relative_path)
adls_output_path= 'abfss://%s@%s.dfs.core.windows.net/%s' % (container_name,
    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/MIrgBebJa+jx7q3PWxqmKNN/H6bGjAvziNrB+
    AStx1PzSA==")

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),
```

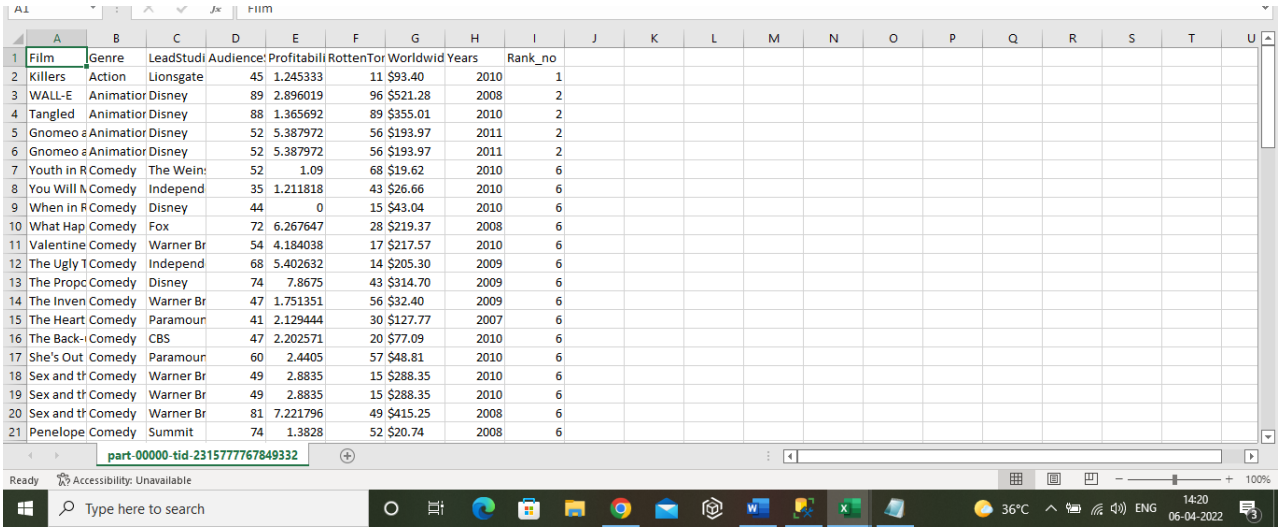
```
    StructField("WorldwideGross",StringType(),True),
    StructField("Years",IntegerType(),True),

])

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 :



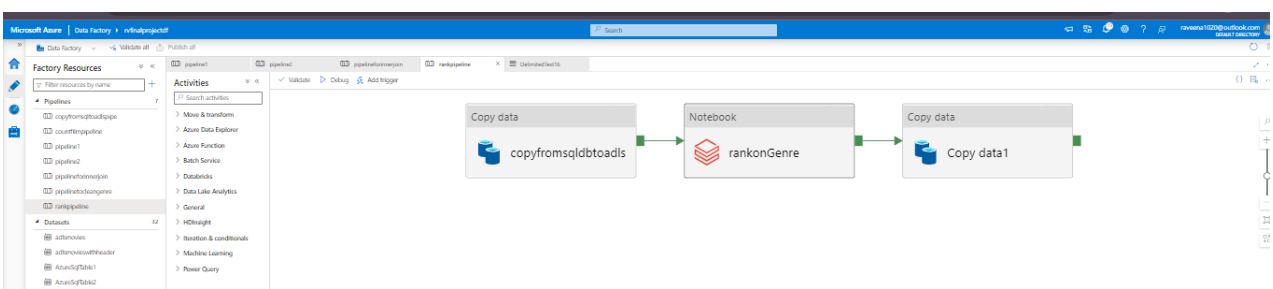


Film	Genre	LeadStudi	Audiences	Profitabili	RottenTor	Worldwid	Years	Rank_no
Killers	Action	Lionsgate	45	1.245333	11	$93.40	2010	1
WALL-E	Animatior	Disney	89	2.896019	96	$521.28	2008	2
Tangled	Animatior	Disney	88	1.365692	89	$355.01	2010	2
Gnomeo a	Animatior	Disney	52	5.387972	56	$193.97	2011	2
Gnomeo a	Animatior	Disney	52	5.387972	56	$193.97	2011	2
Youth in R	Comedy	The Weins	52	1.09	68	$19.62	2010	6
You Will M	Comedy	Independ	35	1.211818	43	$26.66	2010	6
When in R	Comedy	Disney	44	0	15	$43.04	2010	6
What Hap	Comedy	Fox	72	6.267647	28	$219.37	2008	6
Valentine	Comedy	Warner Br	54	4.184038	17	$217.57	2010	6
The Ugly T	Comedy	Independ	68	5.402632	14	$205.30	2009	6
The Propo	Comedy	Disney	74	7.8675	43	$314.70	2009	6
The Inven	Comedy	Warner Br	47	1.751351	56	$32.40	2009	6
The Heart	Comedy	Paramoun	41	2.129444	30	$127.77	2007	6
The Back-(	Comedy	CBS	47	2.202571	20	$77.09	2010	6
She's Out	Comedy	Paramoun	60	2.4405	57	$48.81	2010	6
Sex and th	Comedy	Warner Br	49	2.8835	15	$288.35	2010	6
Sex and th	Comedy	Warner Br	49	2.8835	15	$288.35	2010	6
Sex and th	Comedy	Warner Br	81	7.221796	49	$415.25	2008	6
Penelope	Comedy	Summit	74	1.3828	52	$20.74	2008	6

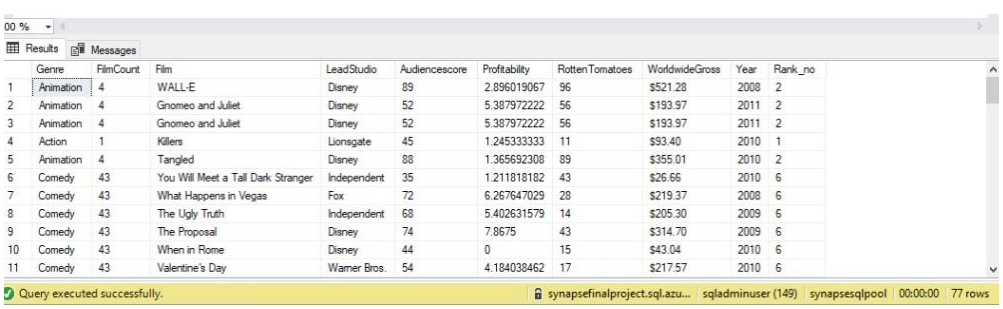
3. Create a final activity that will read the output of previous activity in ADLS and store the same inSynapse.

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

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