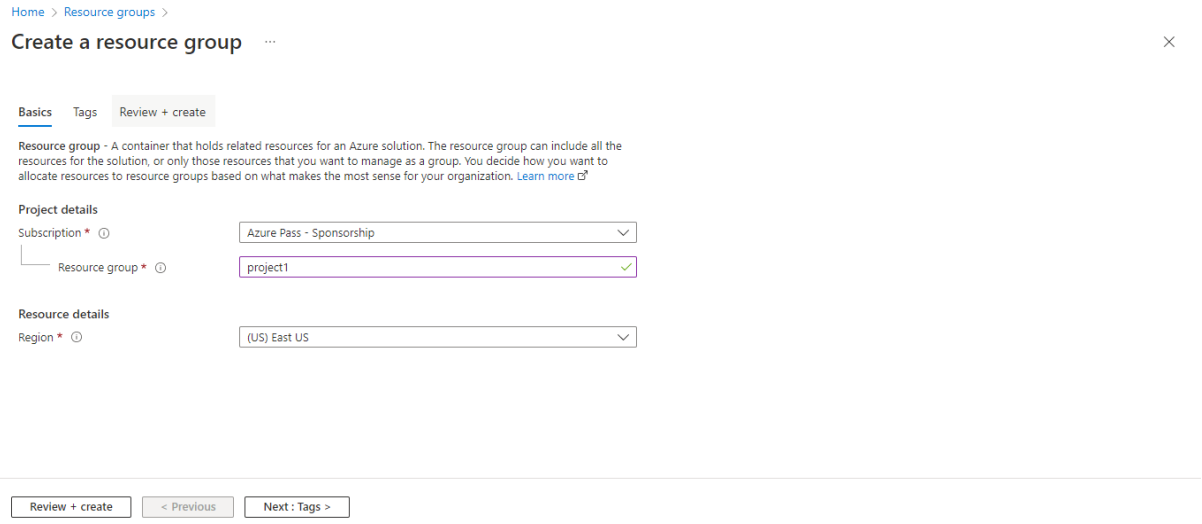
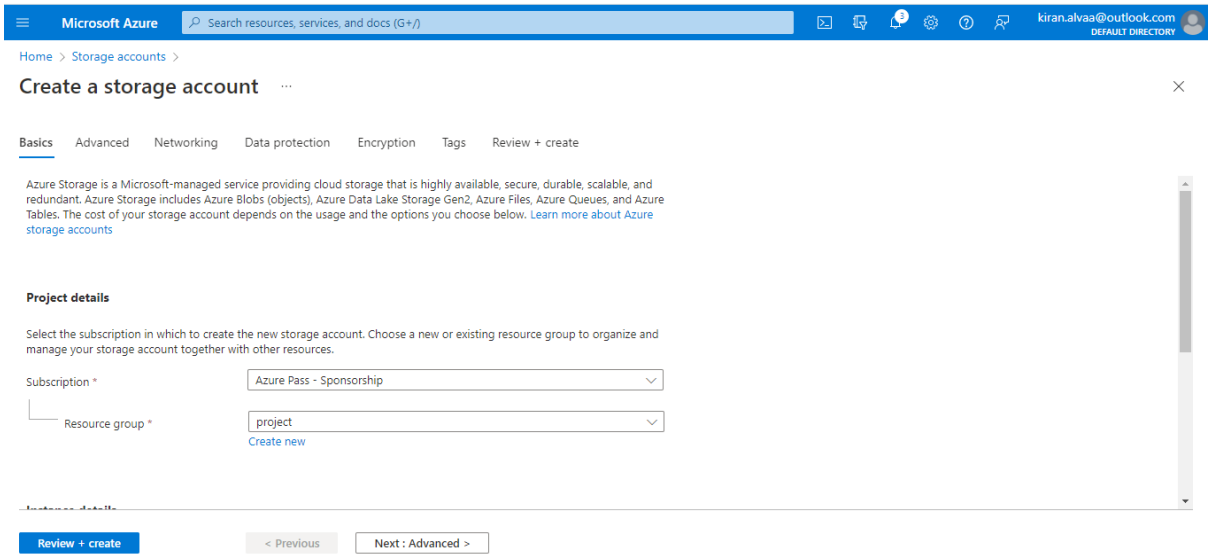
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

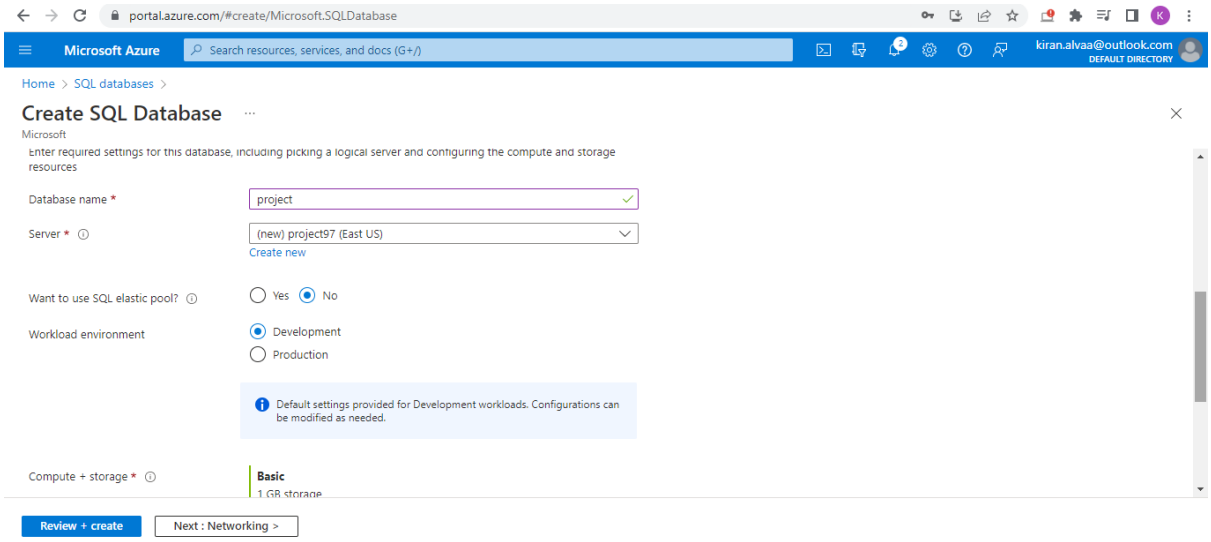
DESCRIPTION Use the data analytics stack to build a data pipeline using Data Factory, Databricks and Synapse. Problem Statement: As a Data Engineer, you've been asked to access the services that can help with ETL of data in the cloud data storage to enable analytics through Synapse. In this POC, we will 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

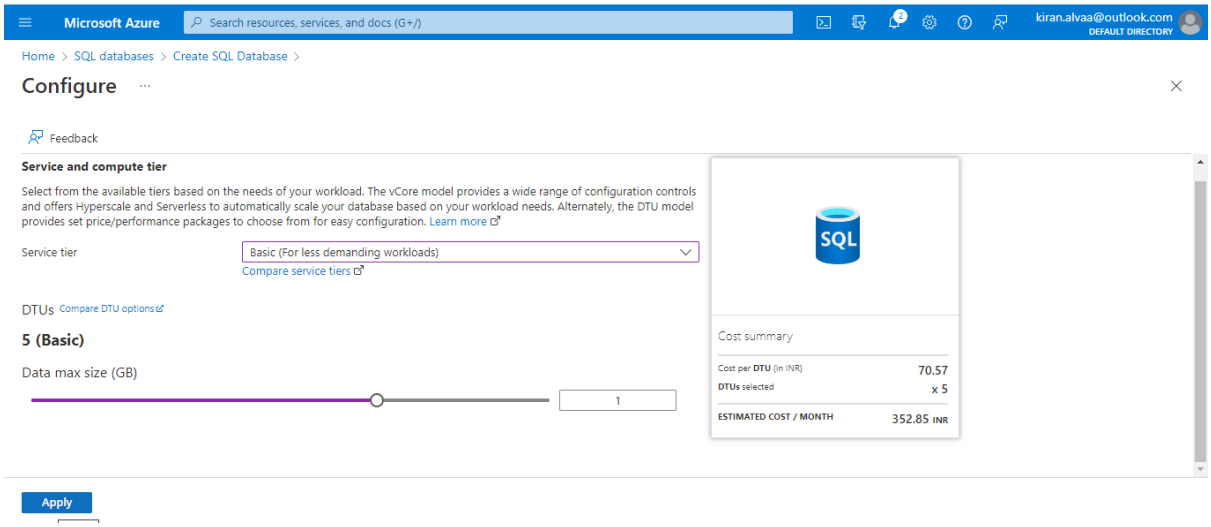
● Create a Resource Group. ● Create a Storage account. ● Create an Azure SQL Database. ● Create a data factory. ● Configure Databricks cluster ● Create Synapse analytics Data Warehouse



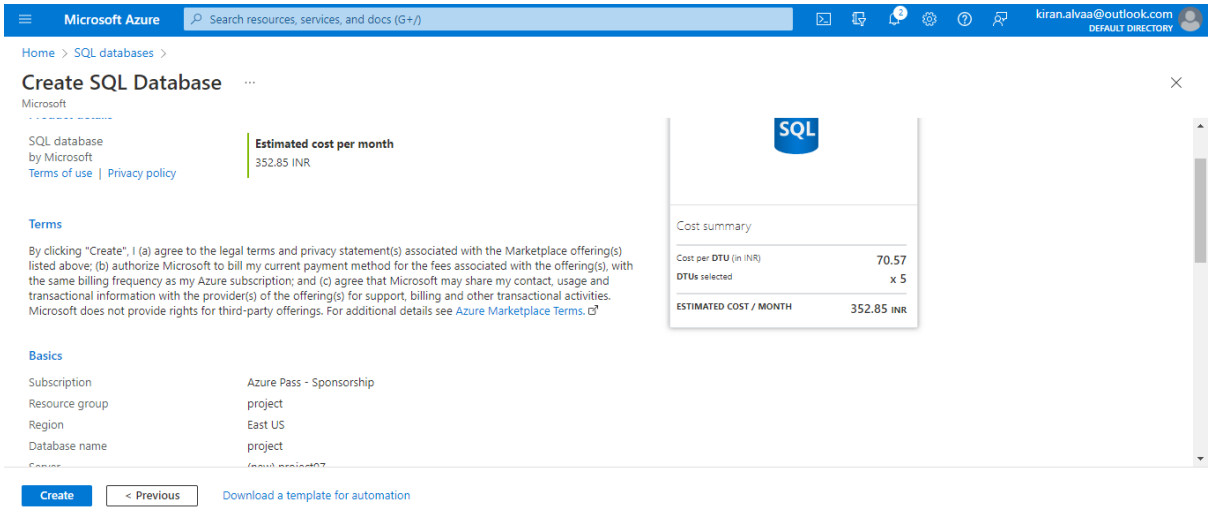


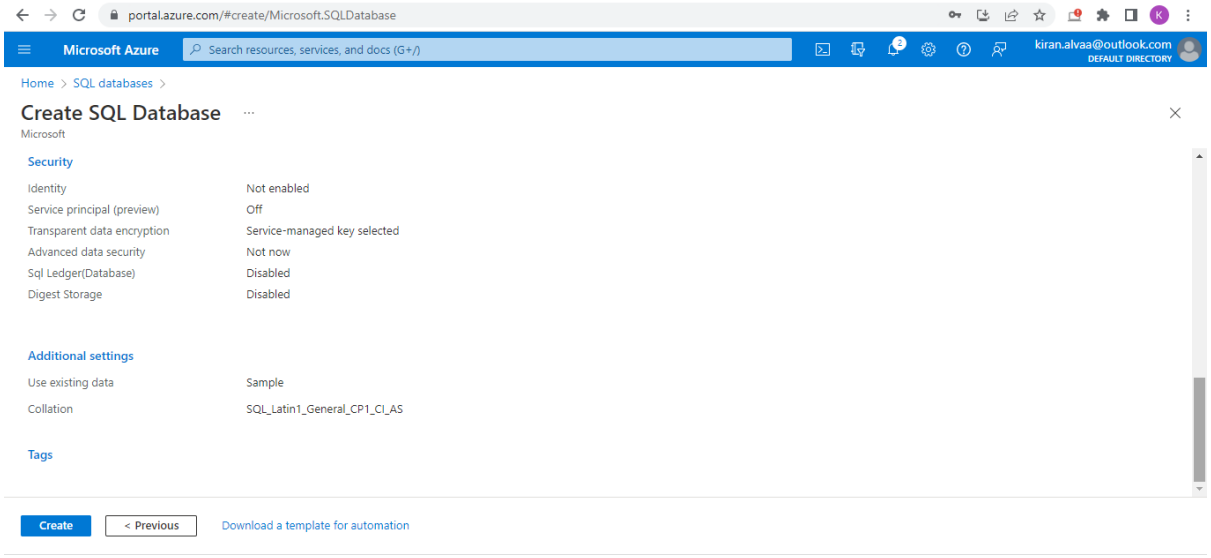
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

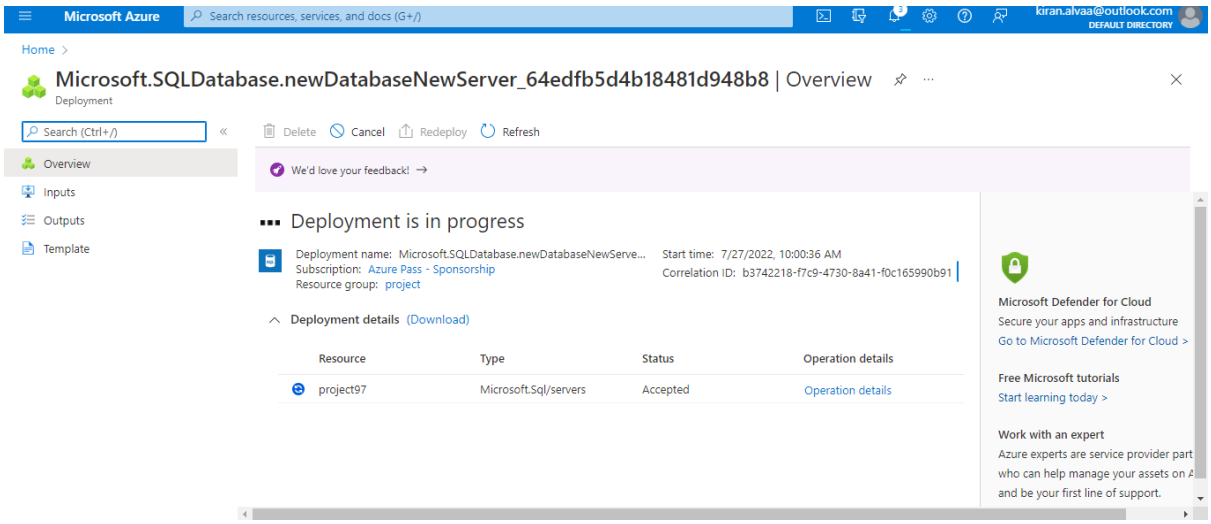




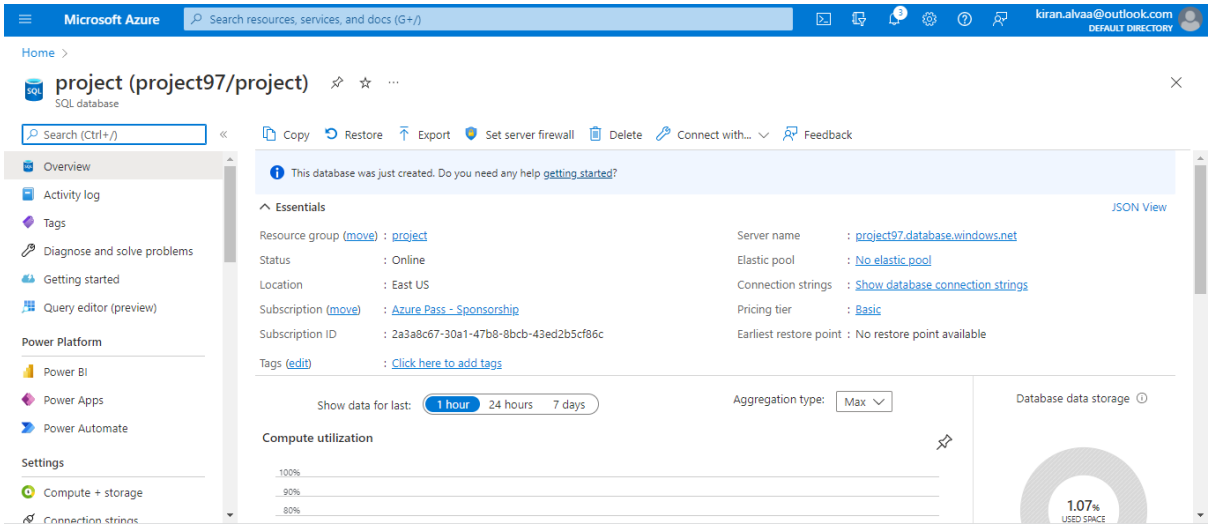
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

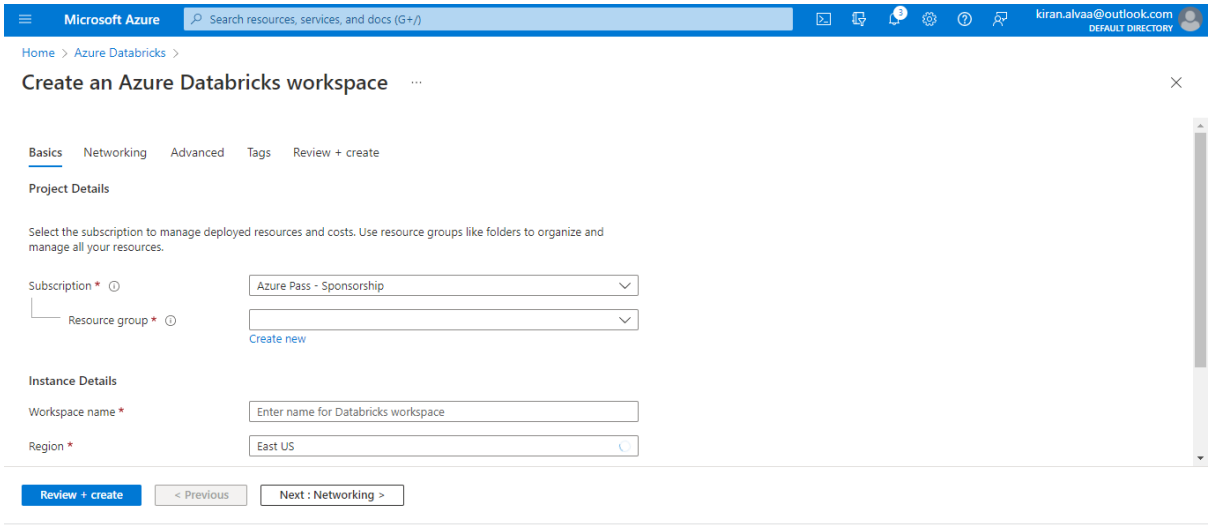




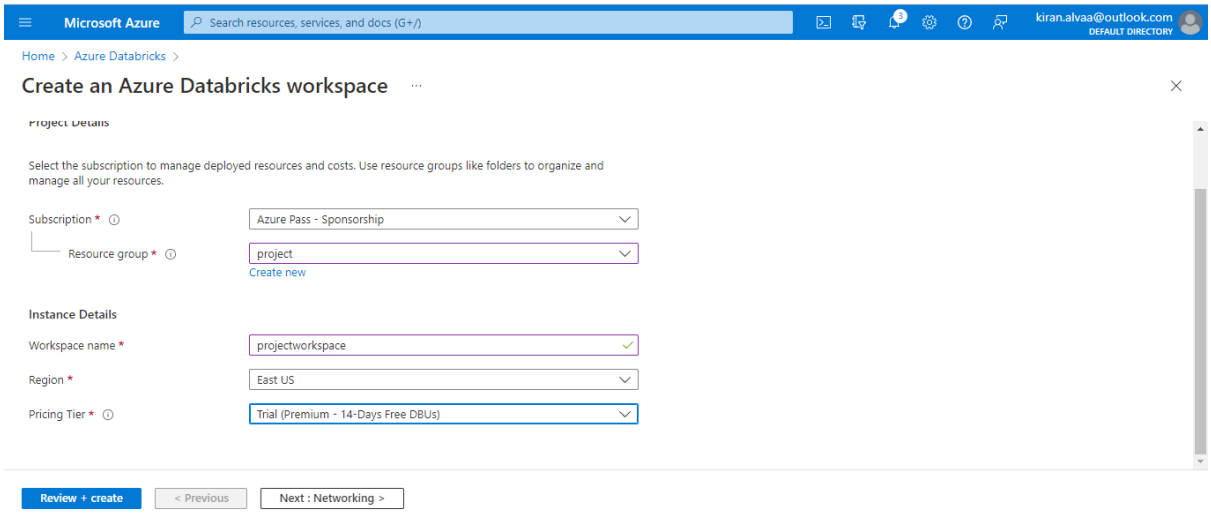


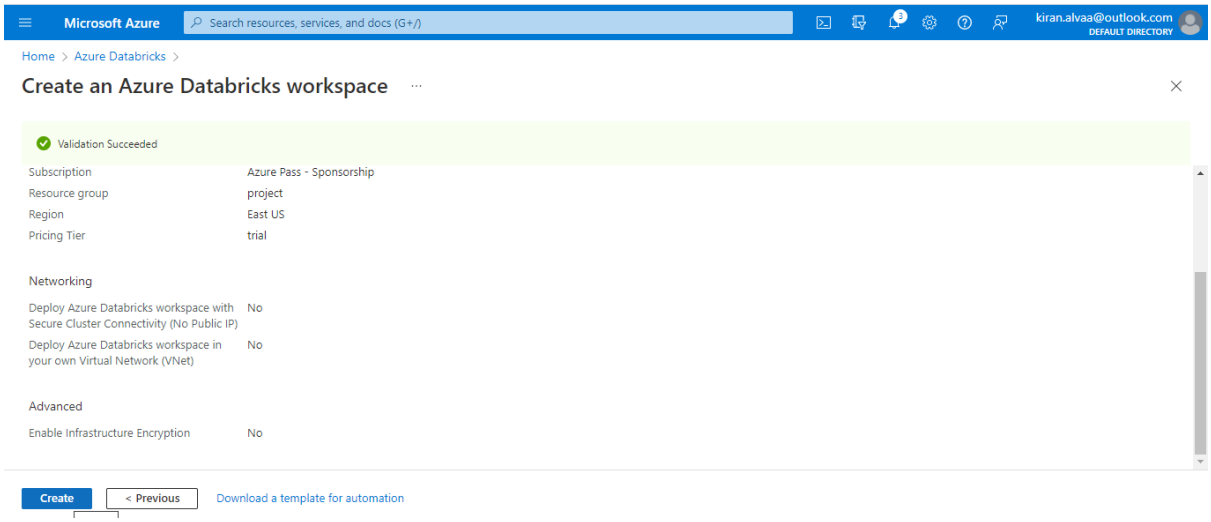
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

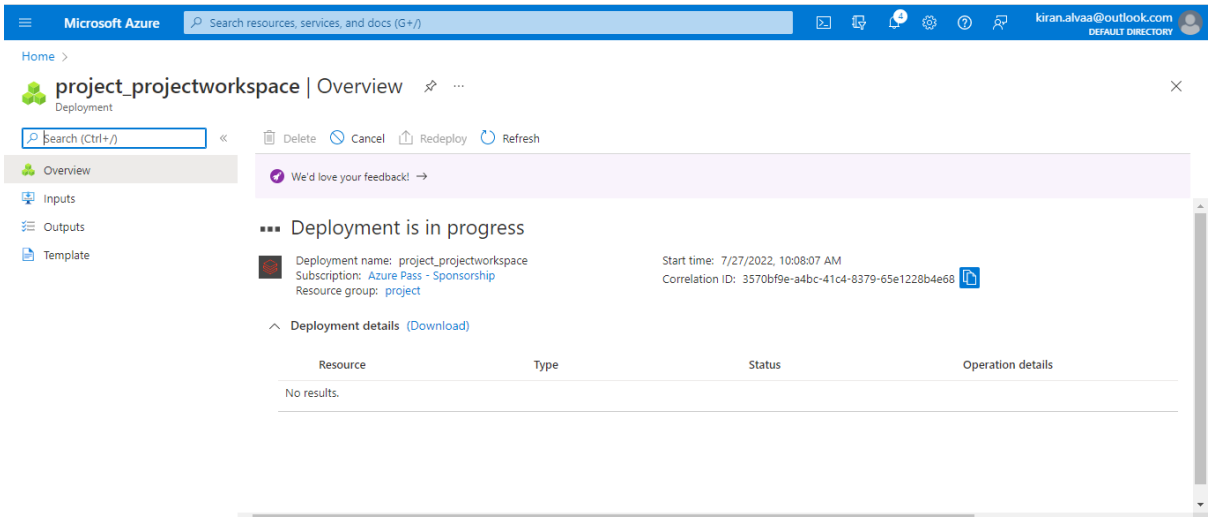




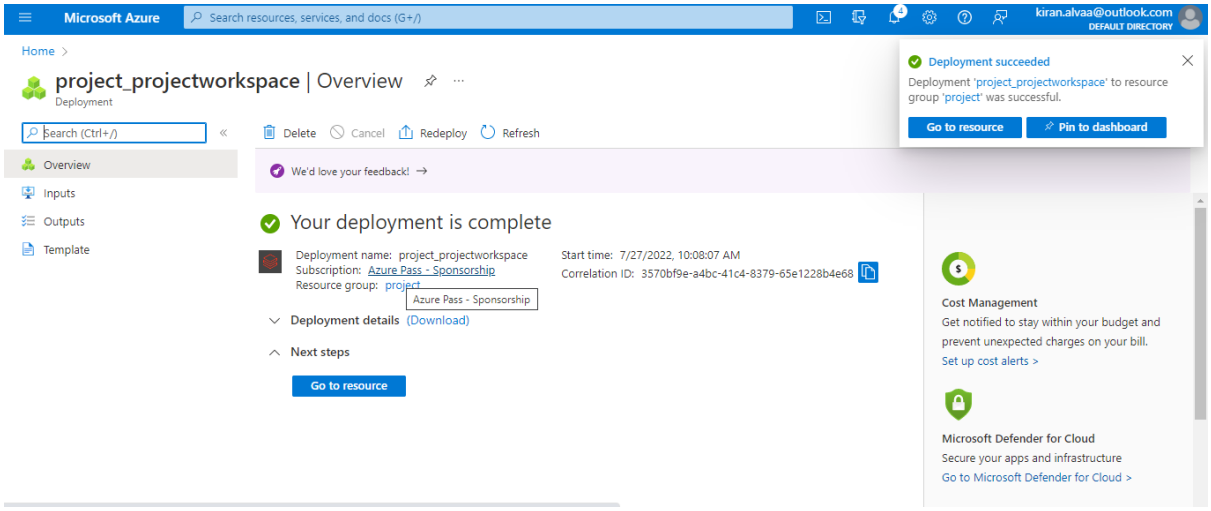
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

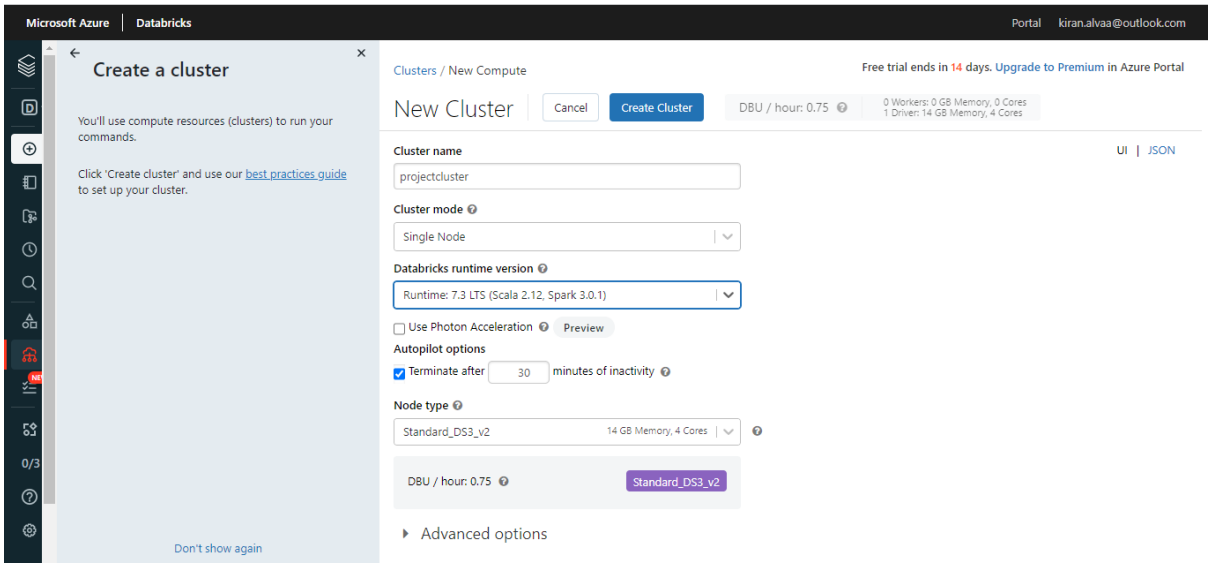


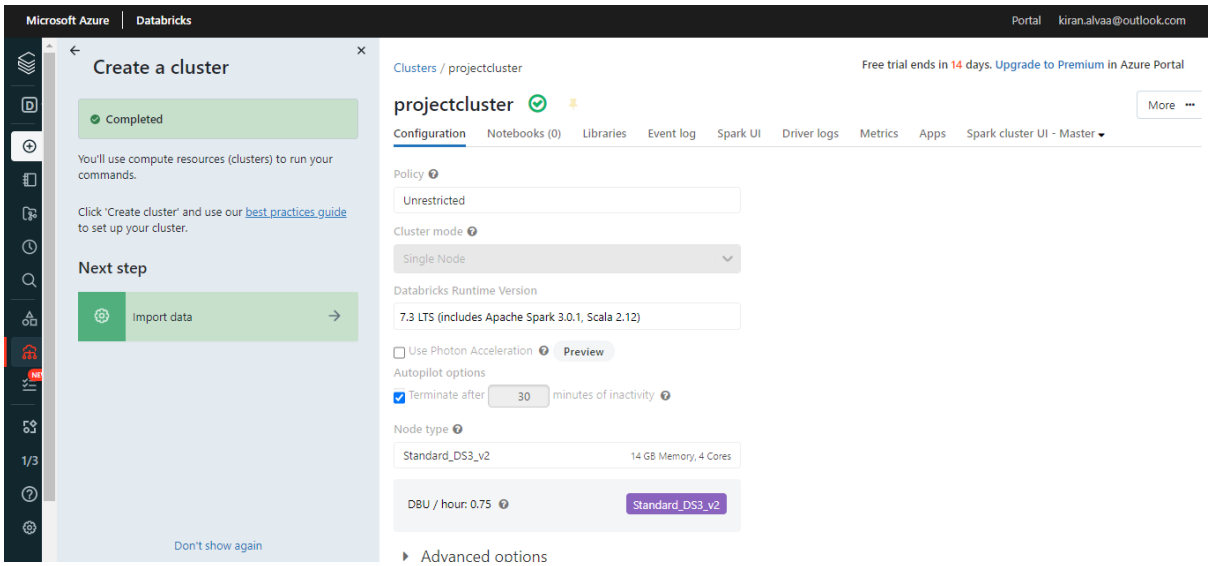




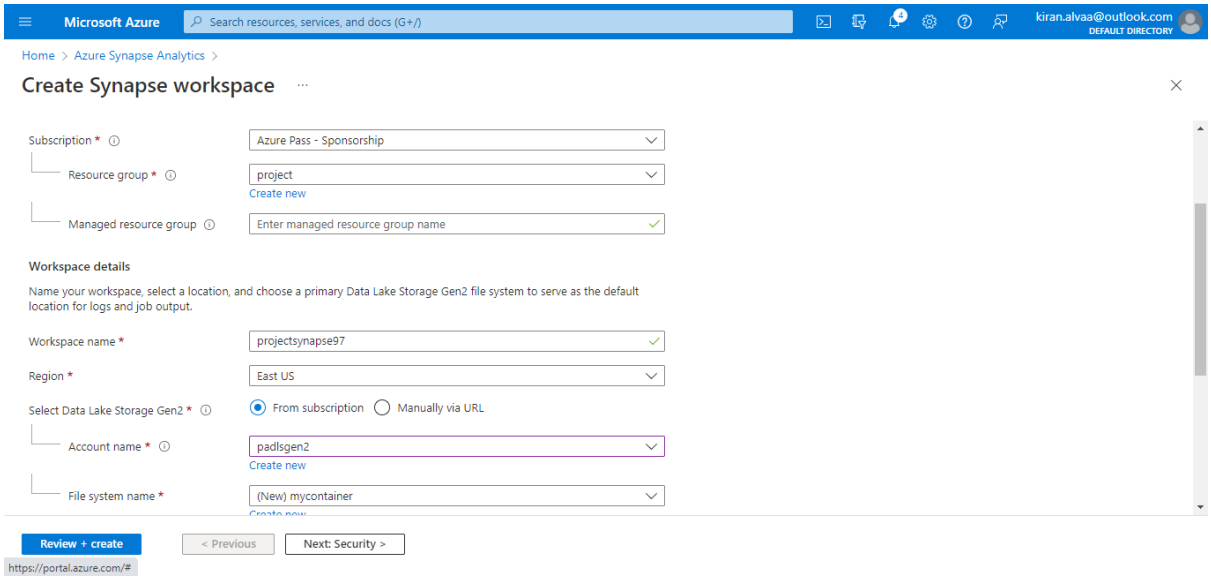
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

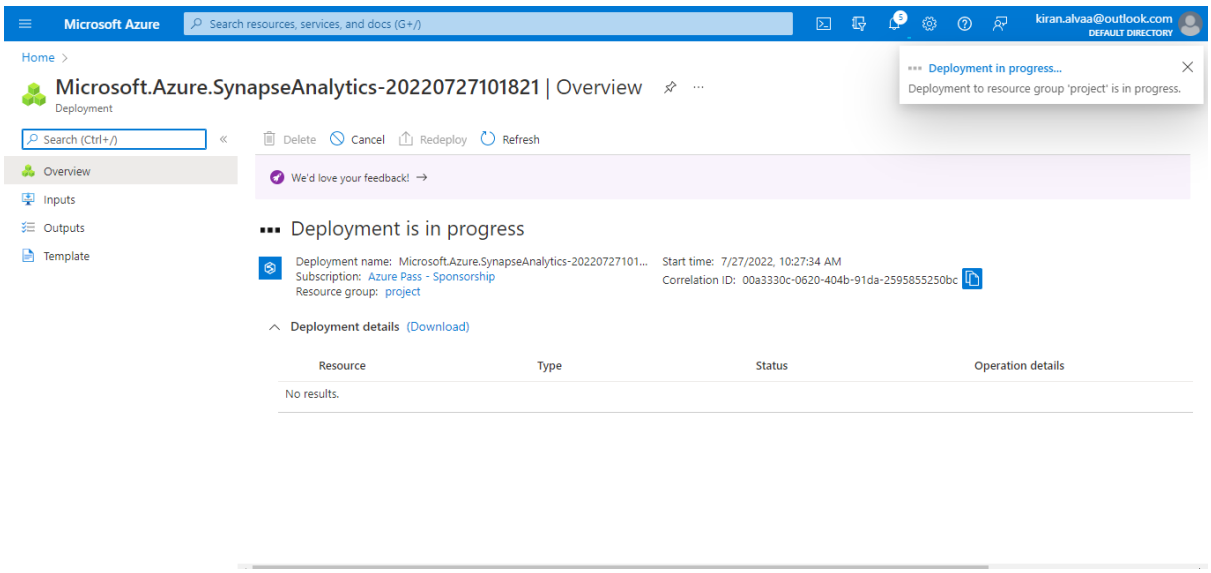




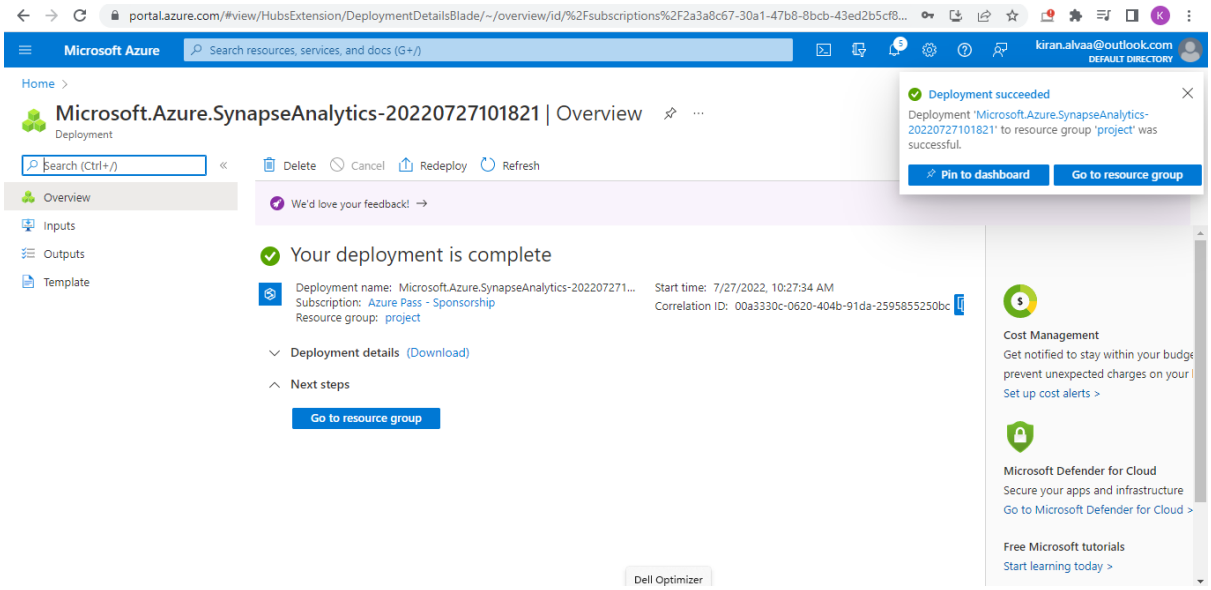


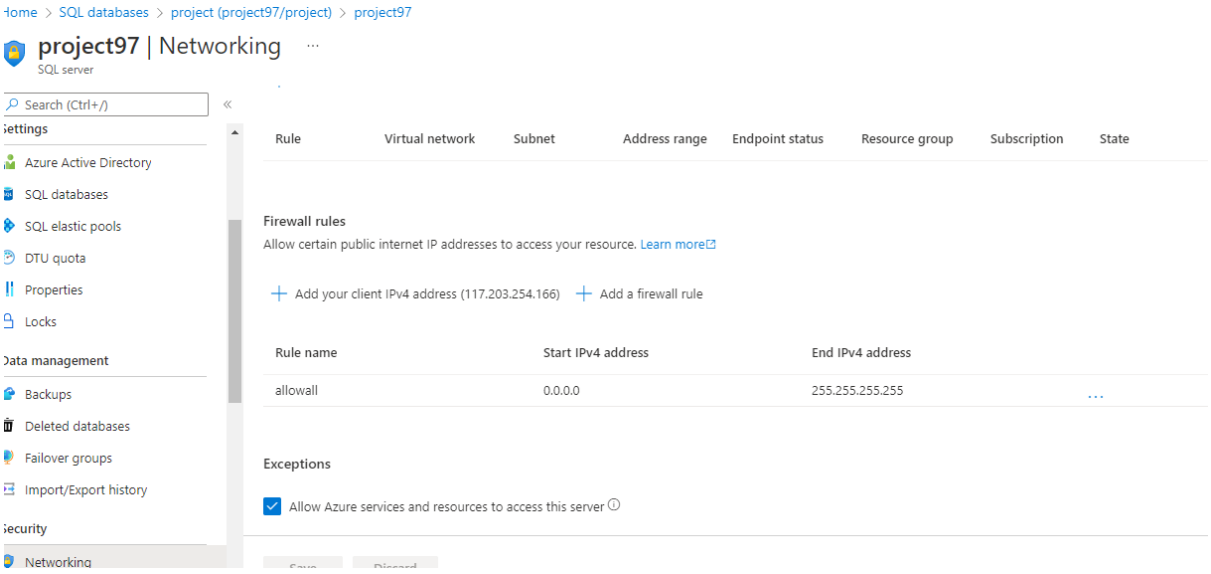
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



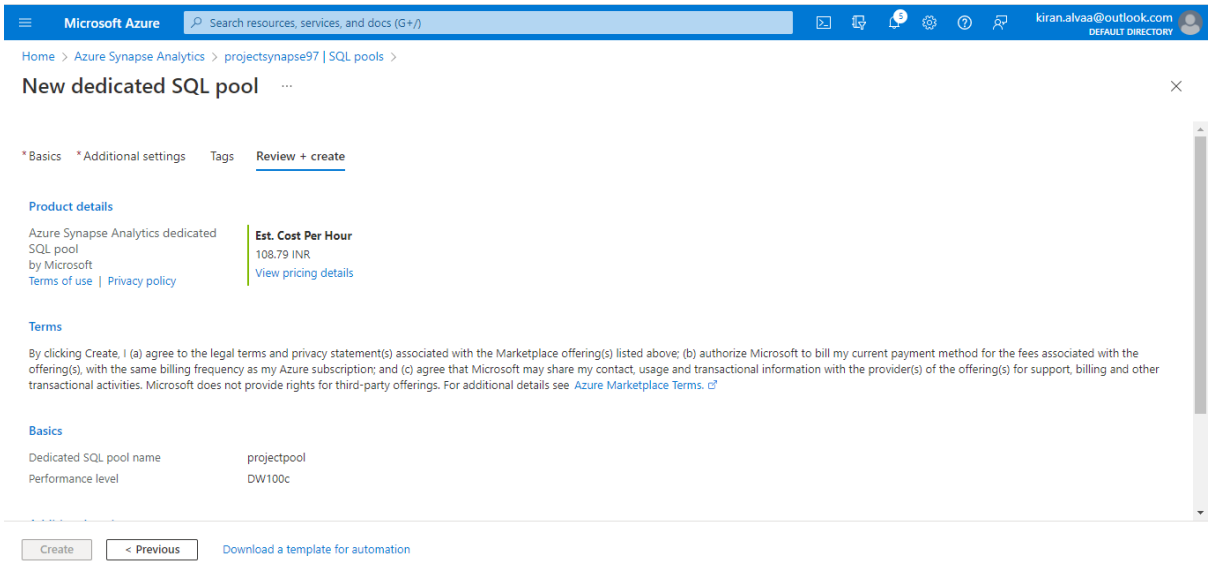


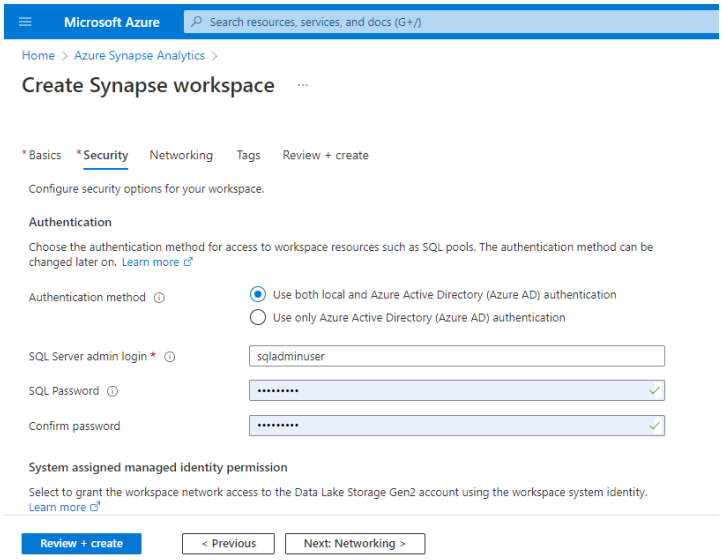
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



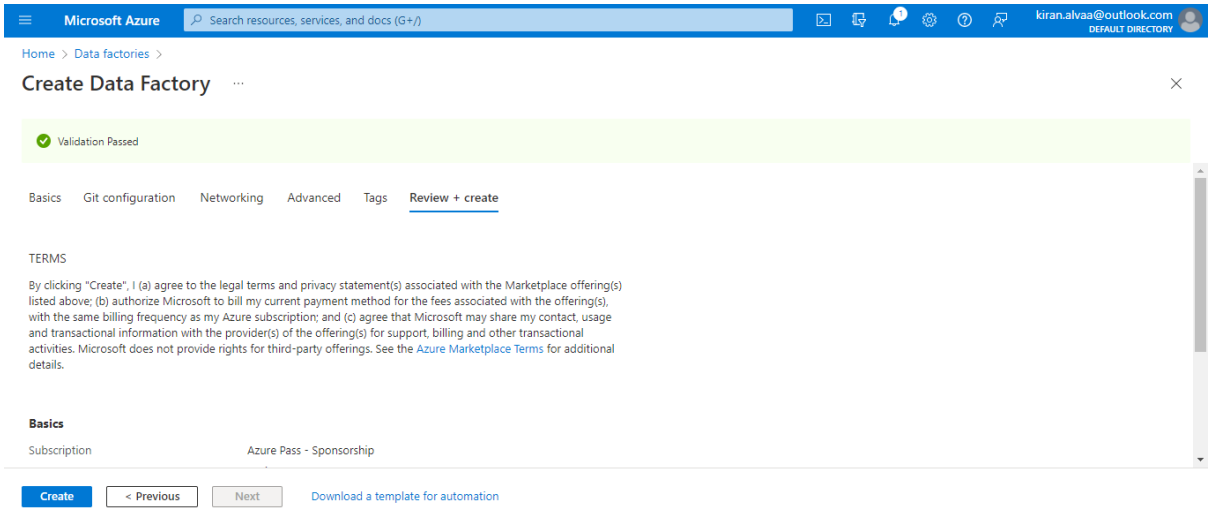


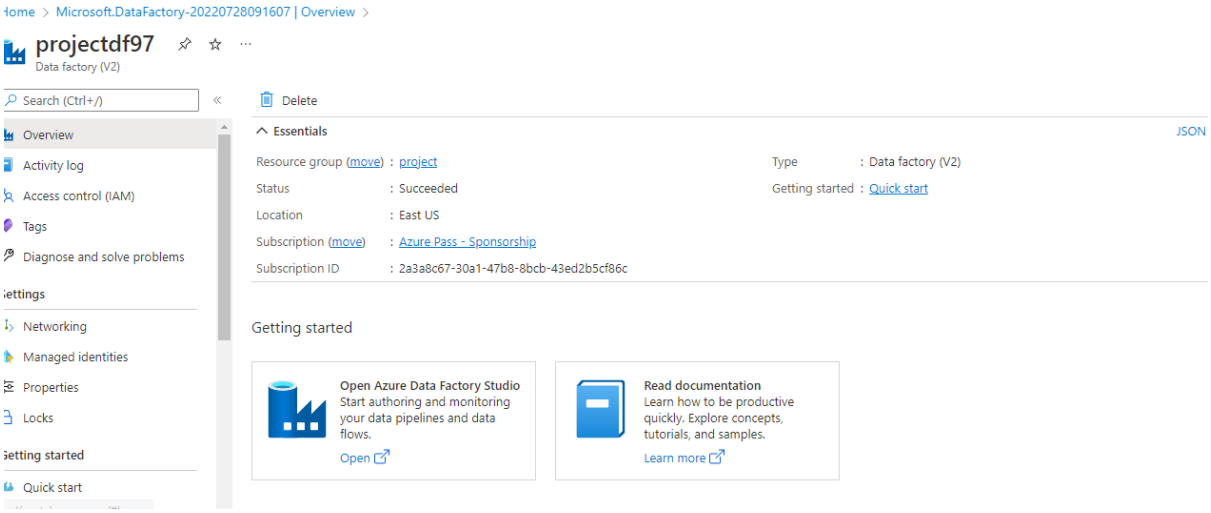
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



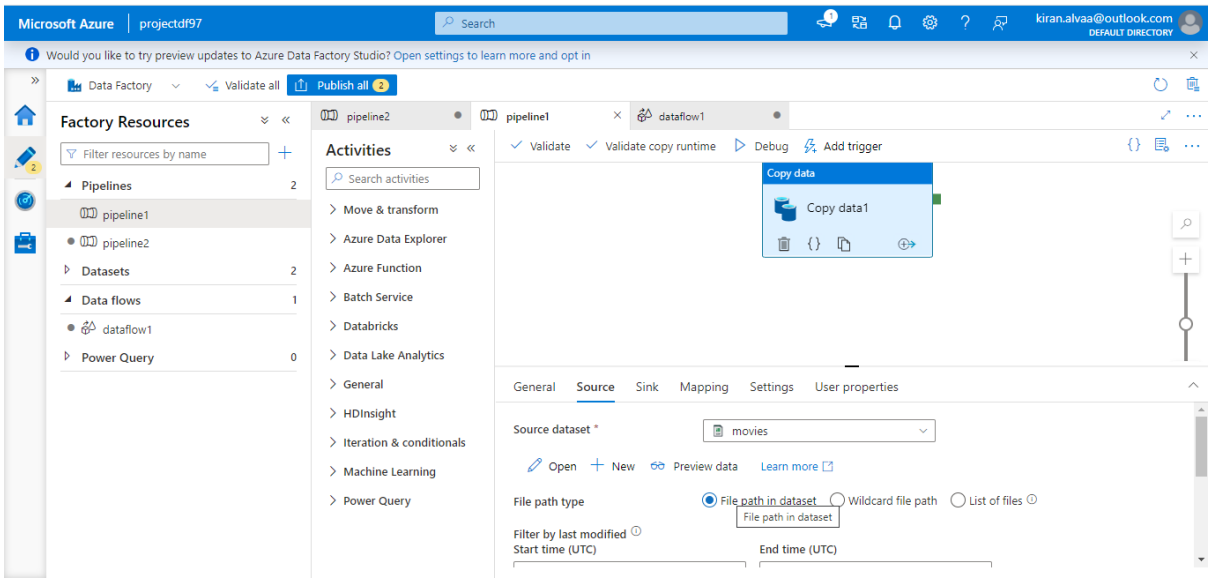


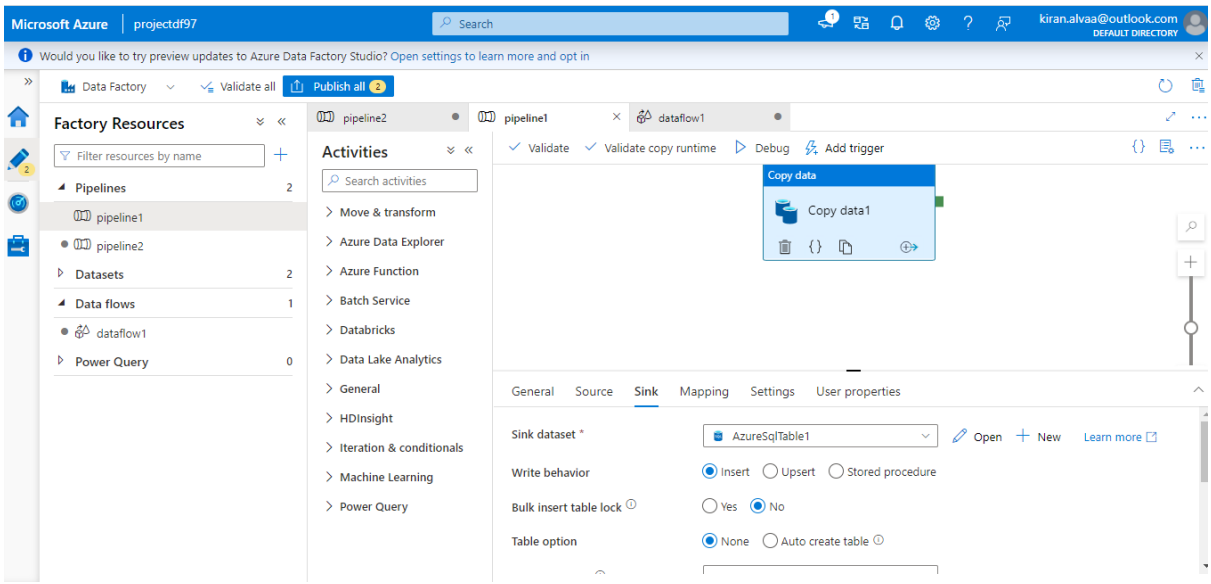
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



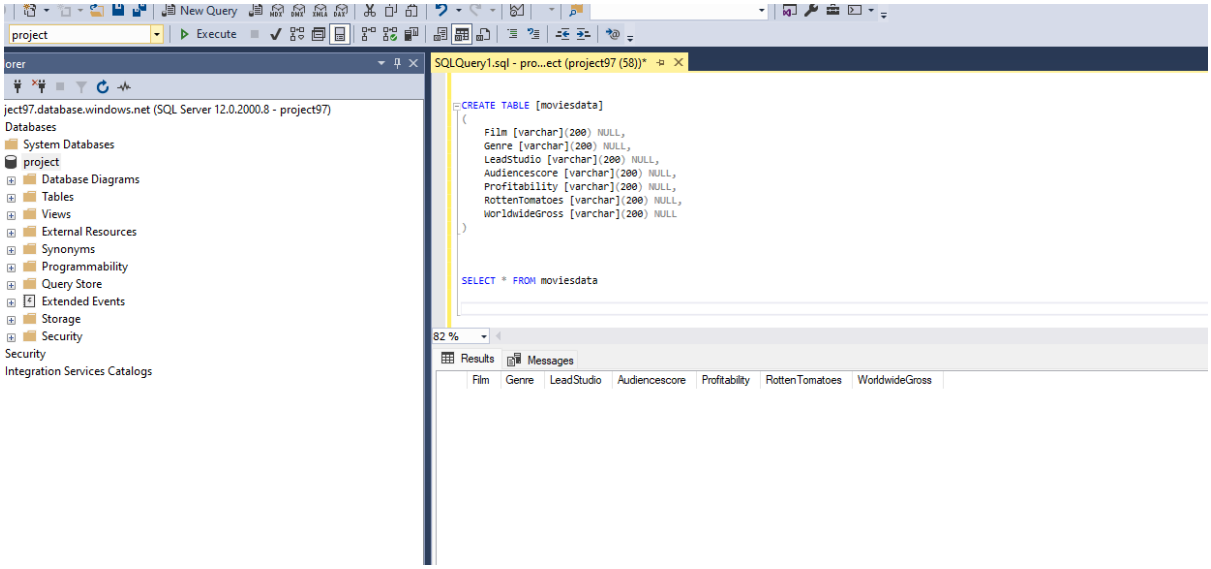


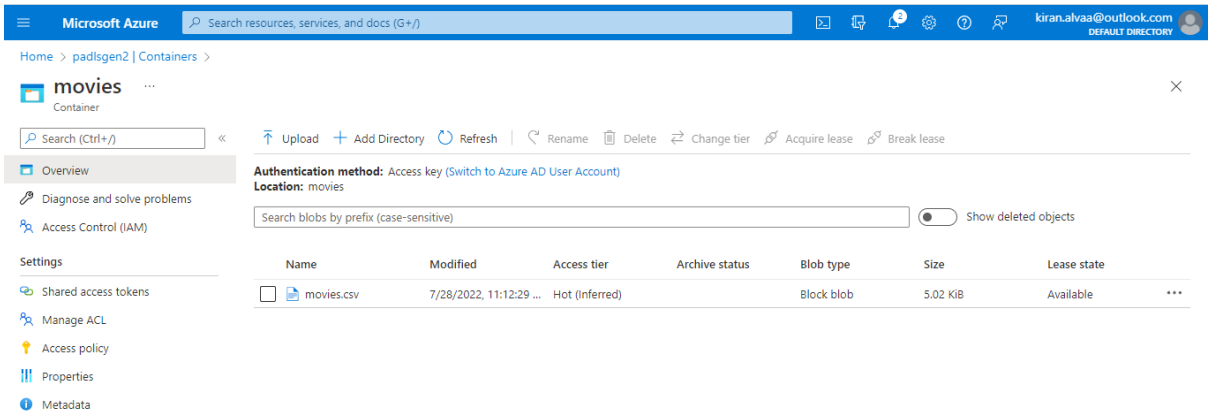
PIPELINE 1 – COPY MOVIES.CSV FROM ADLS TO SQL TABLE

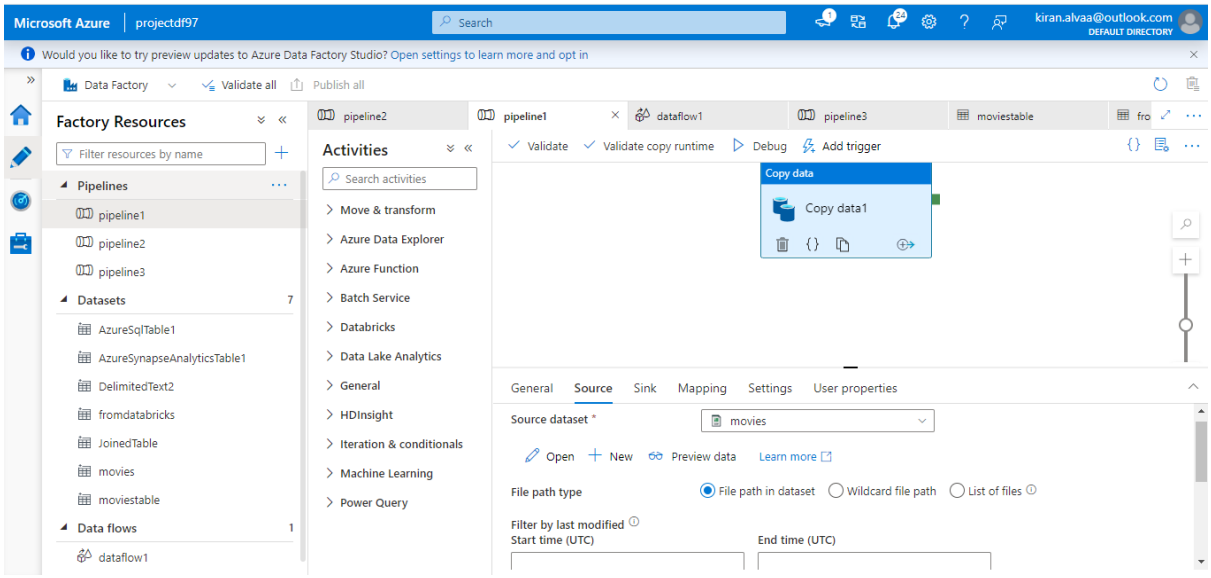




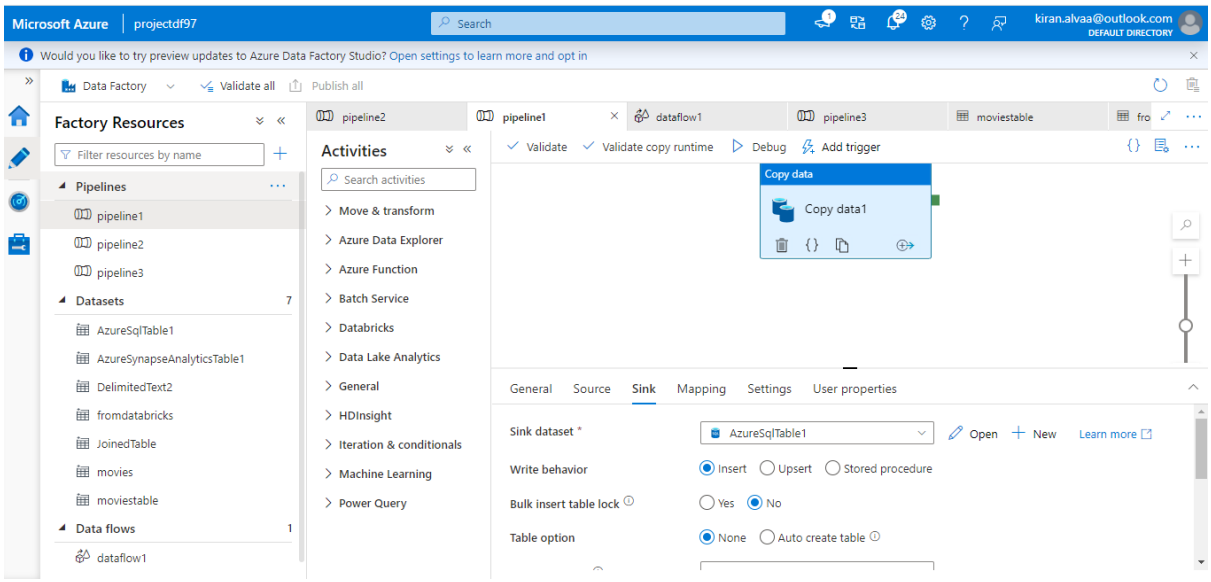
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

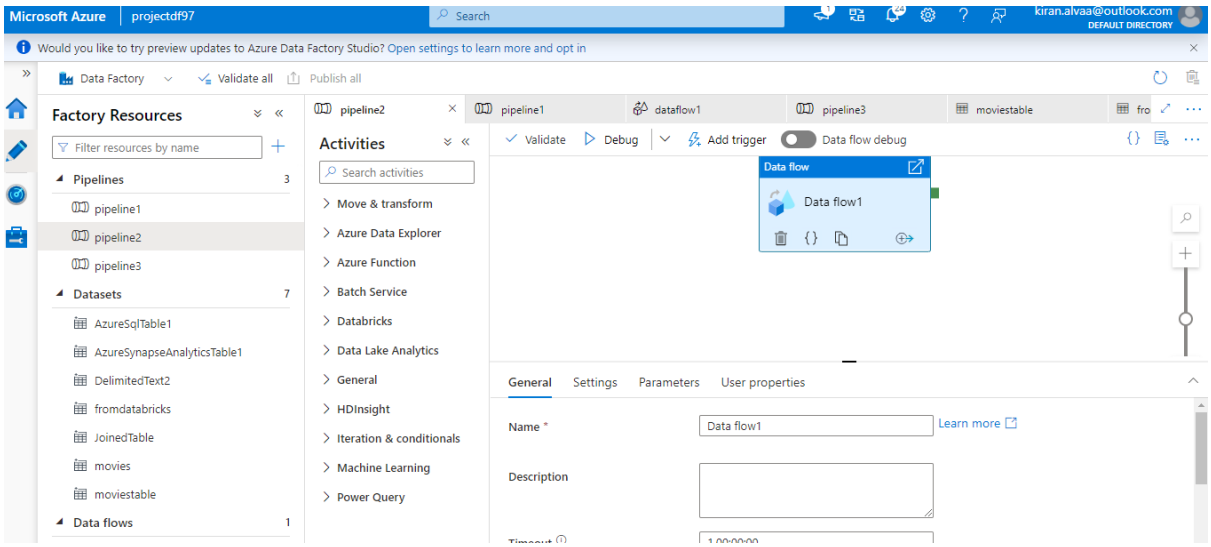






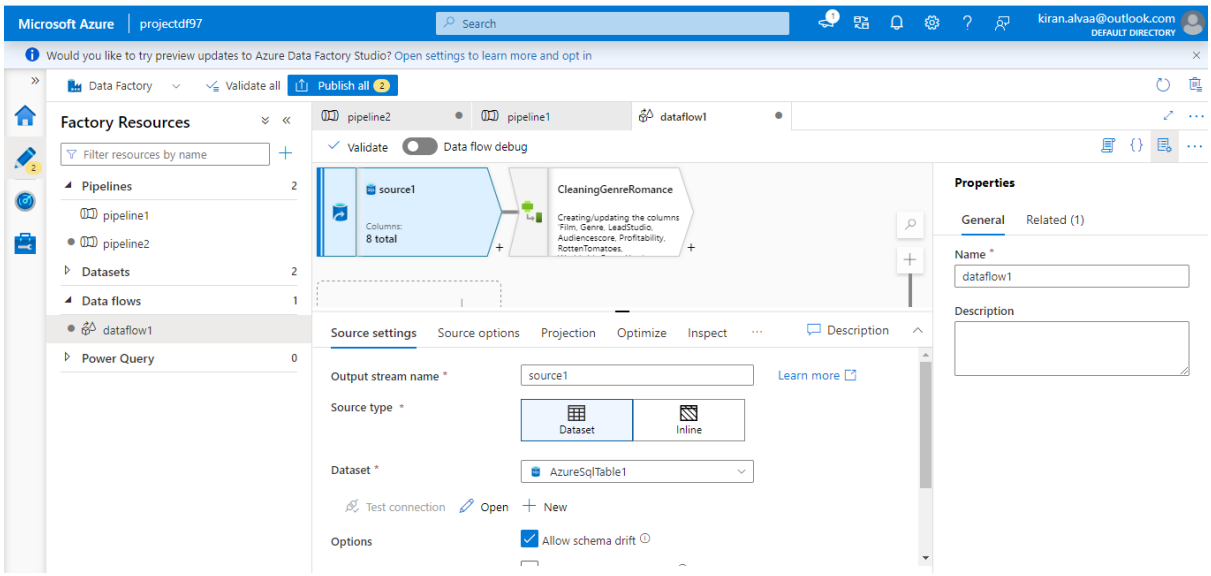
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

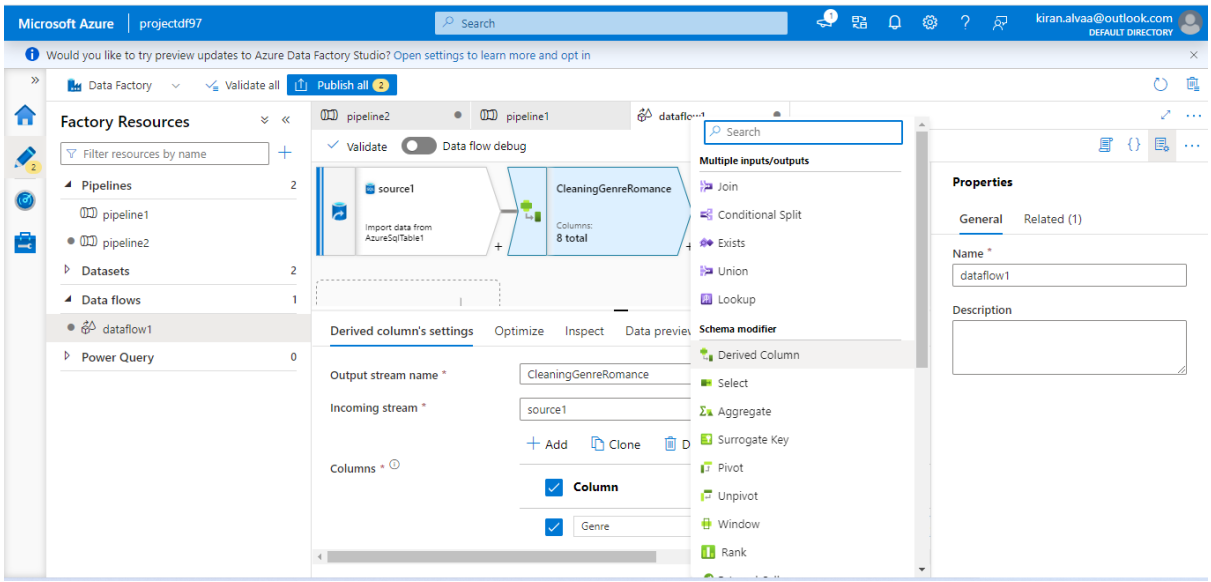


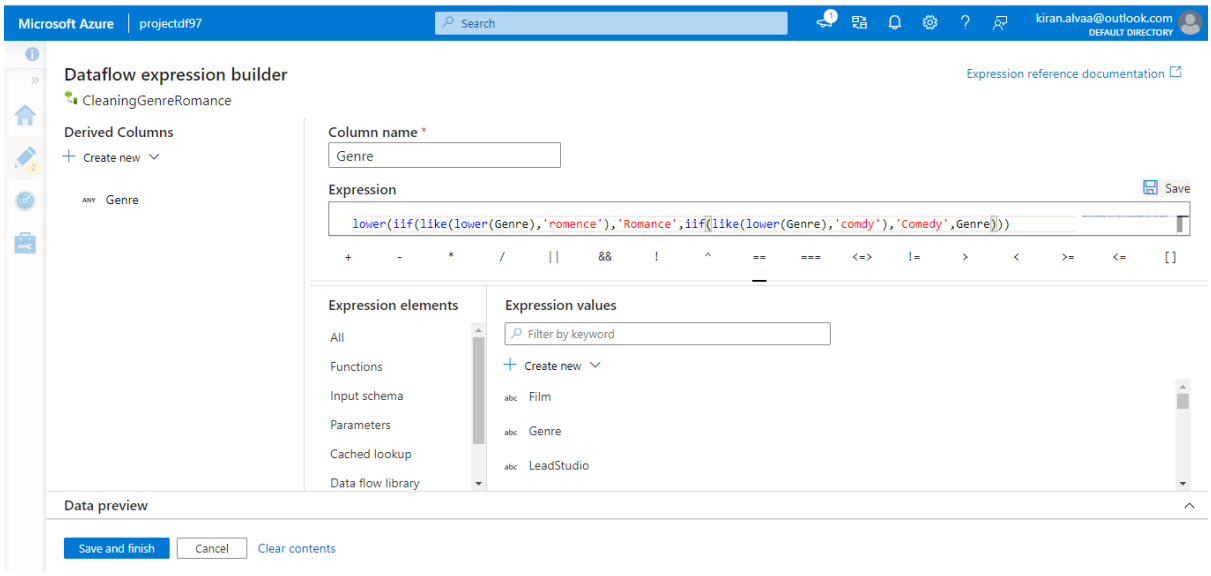


1. Create a data stream named CleaningGenreRomance and perform data cleansing on the Genre column using Derived Column and case expression. (While collecting data it was observed that some genres have spelling mistakes like romance, Romence for Romance, comedy, Comdy for Comedy.)

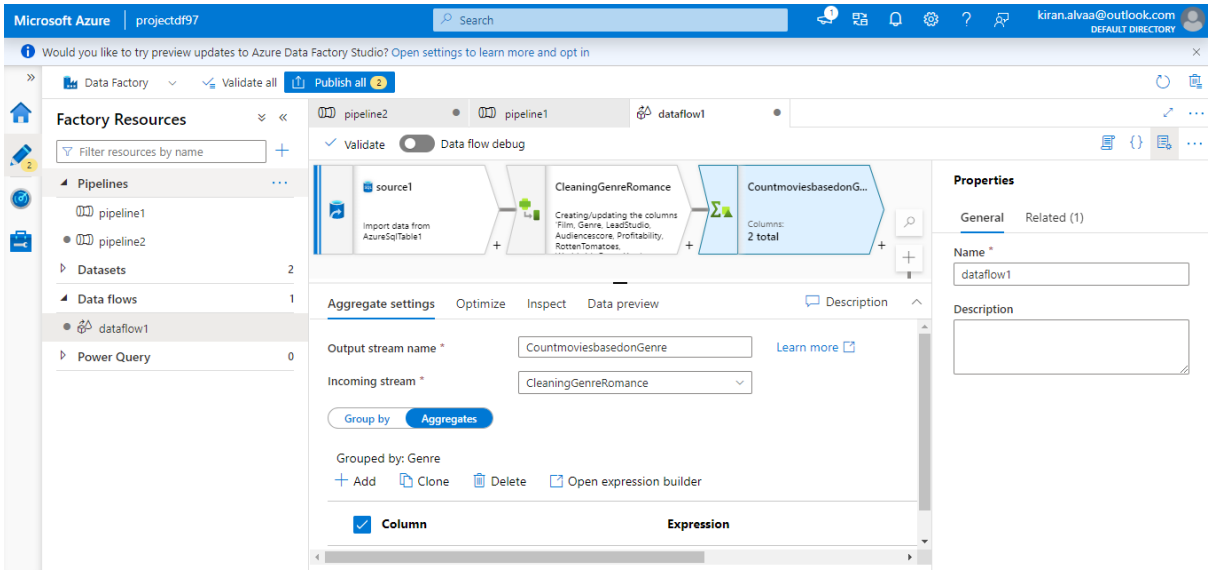
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



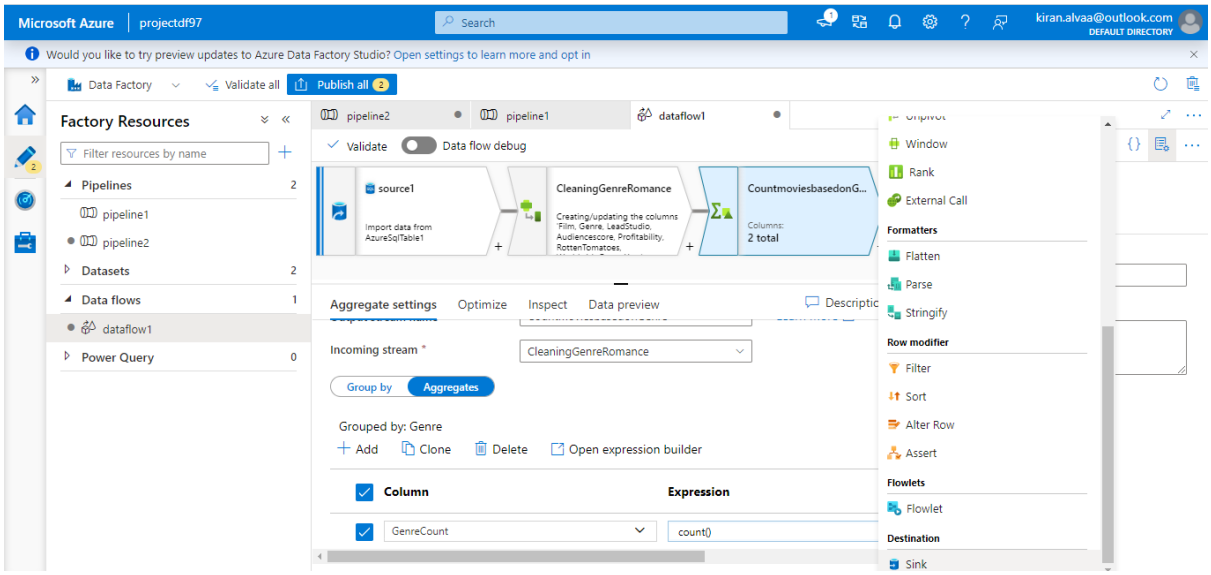


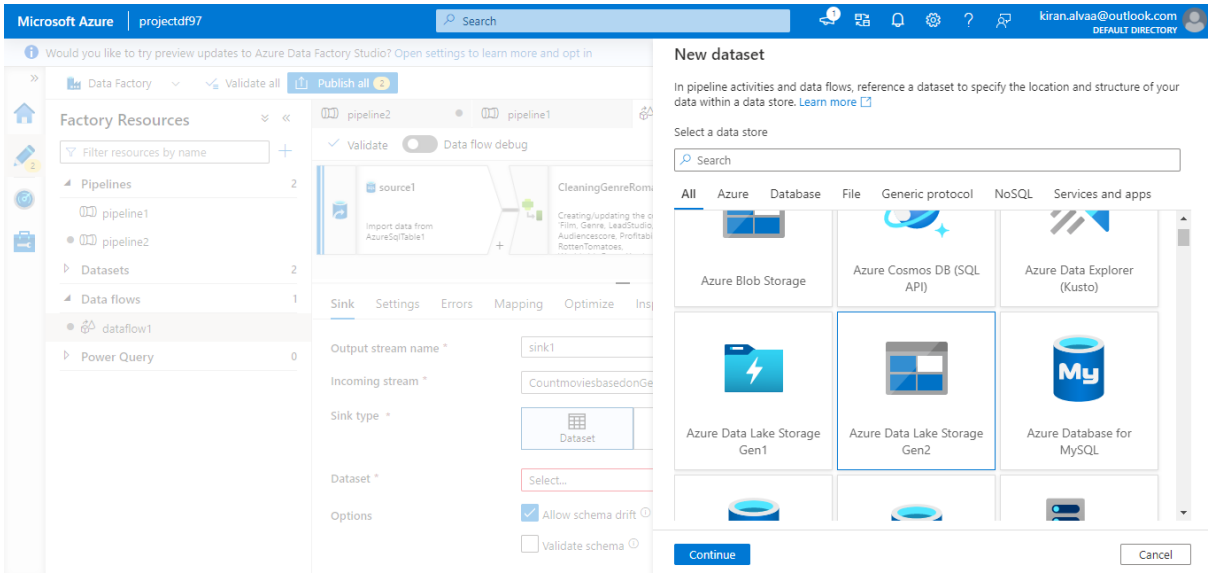


2. Create a data stream named CountMoviesBasedOnGenre that can calculate number of films for each genre and store it as a separate dataset in ADLS under folder name “solution/genreCount”

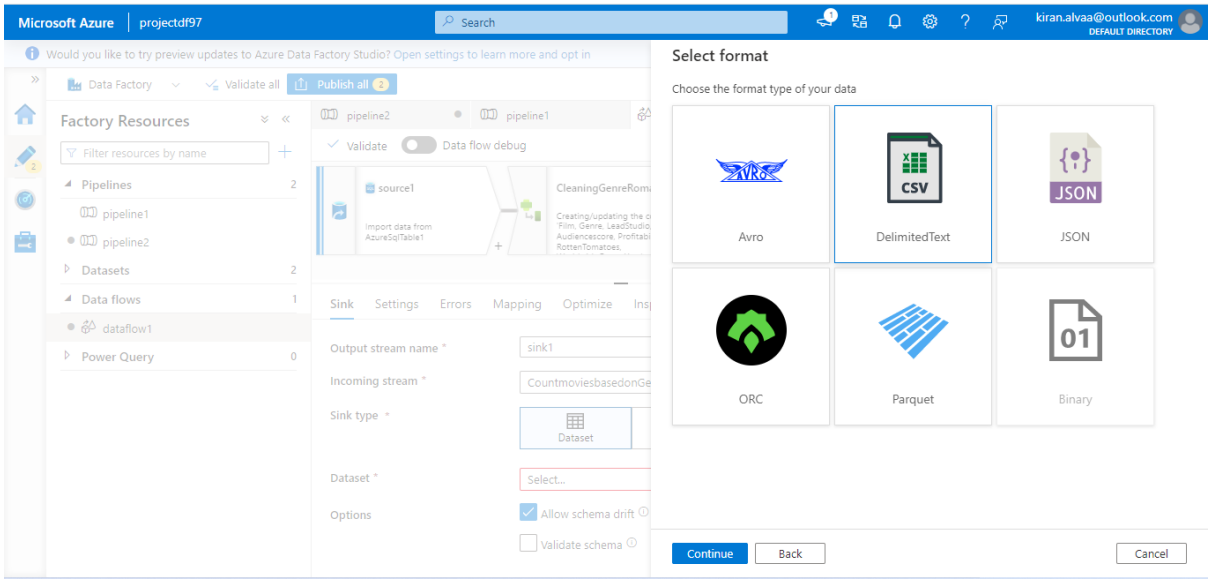


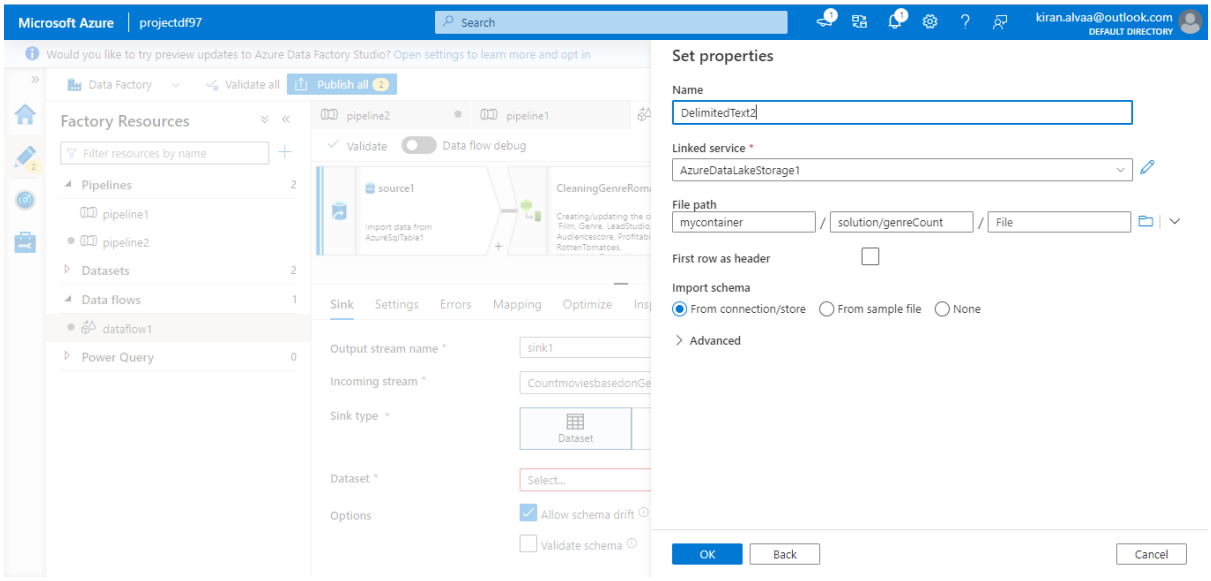
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

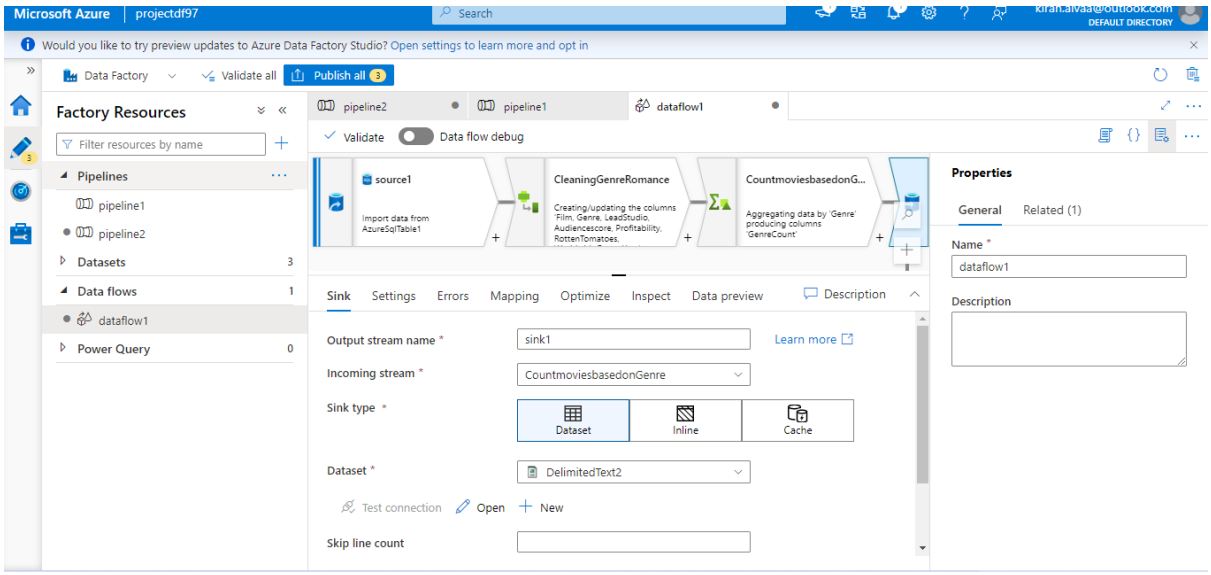




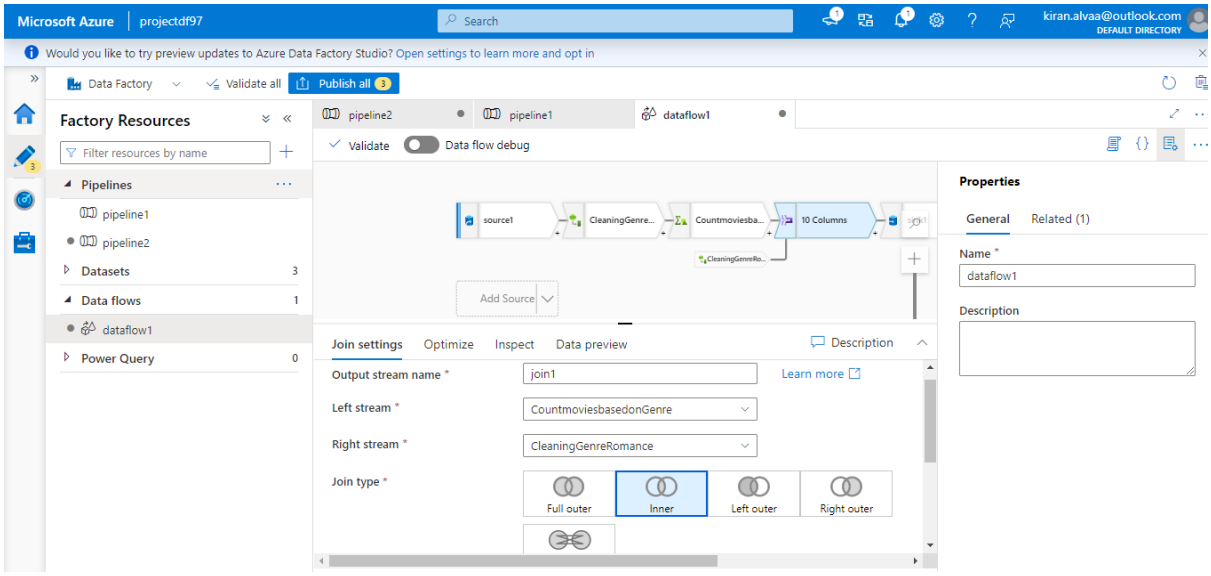
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



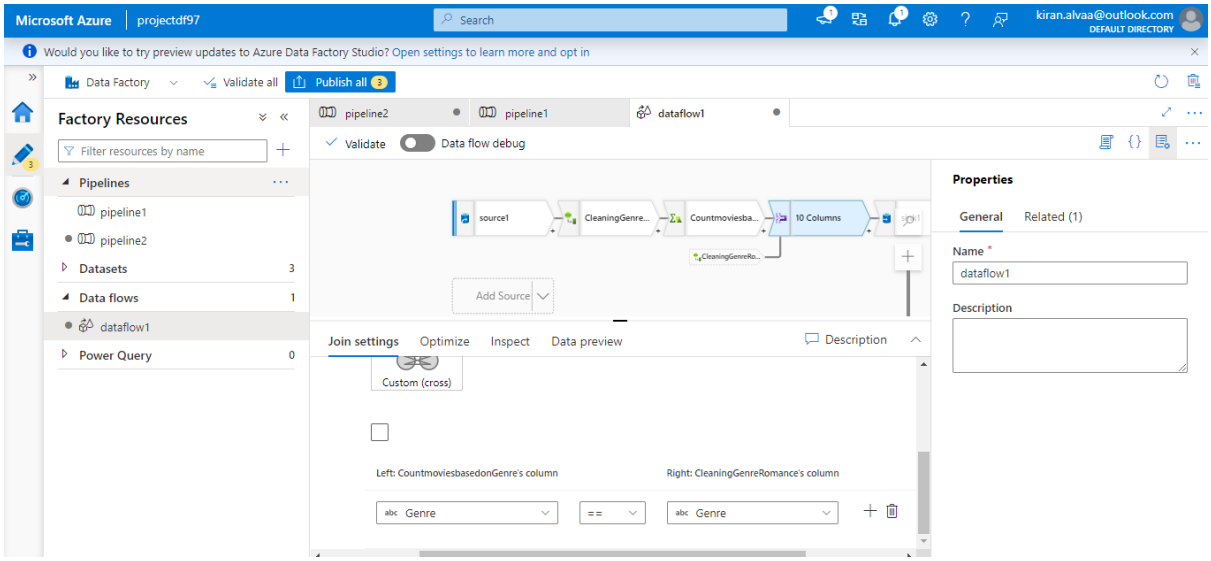


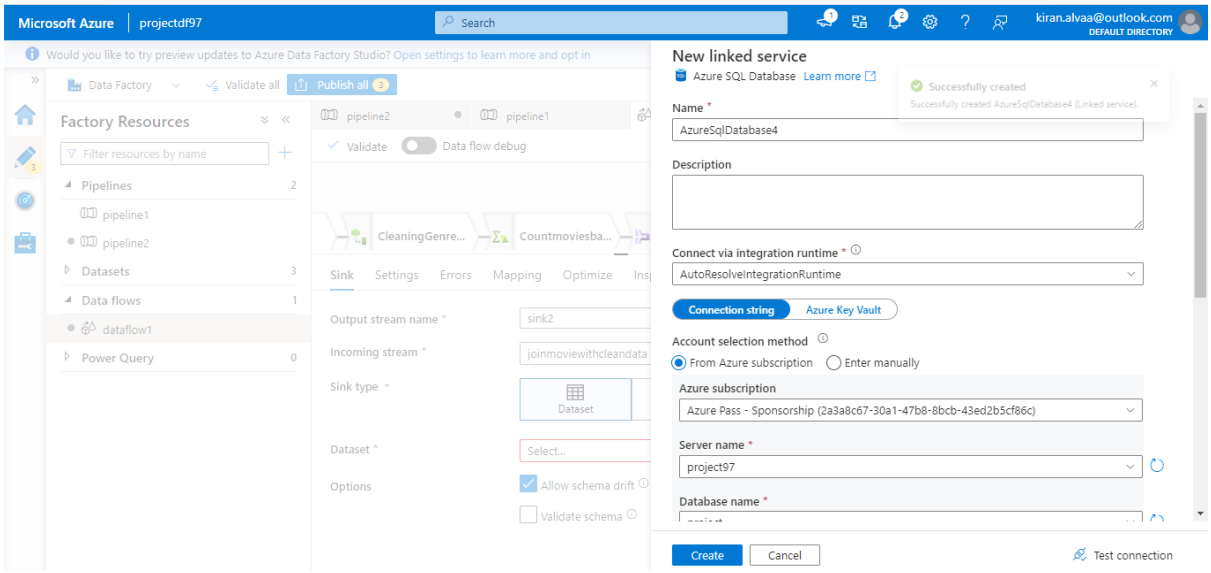


3. Create a new stream named JoinMovieCountWithCleanData. Perform join operation on CountMoviesBasedOnGenre with CleaningGenreRomance stream and store the same in the Azure SQL Database.

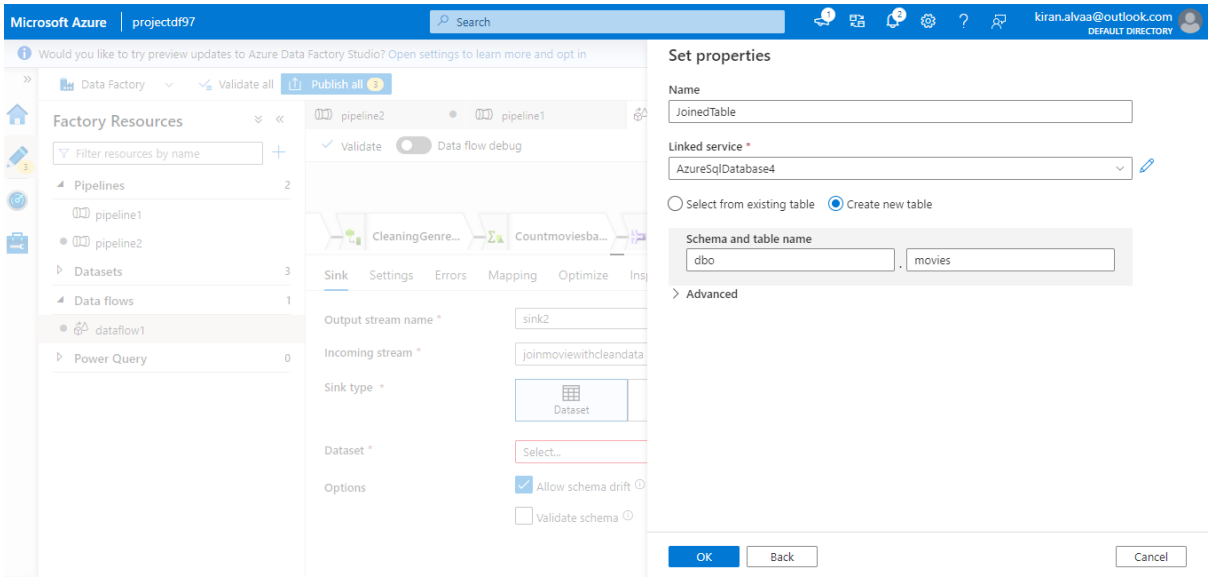


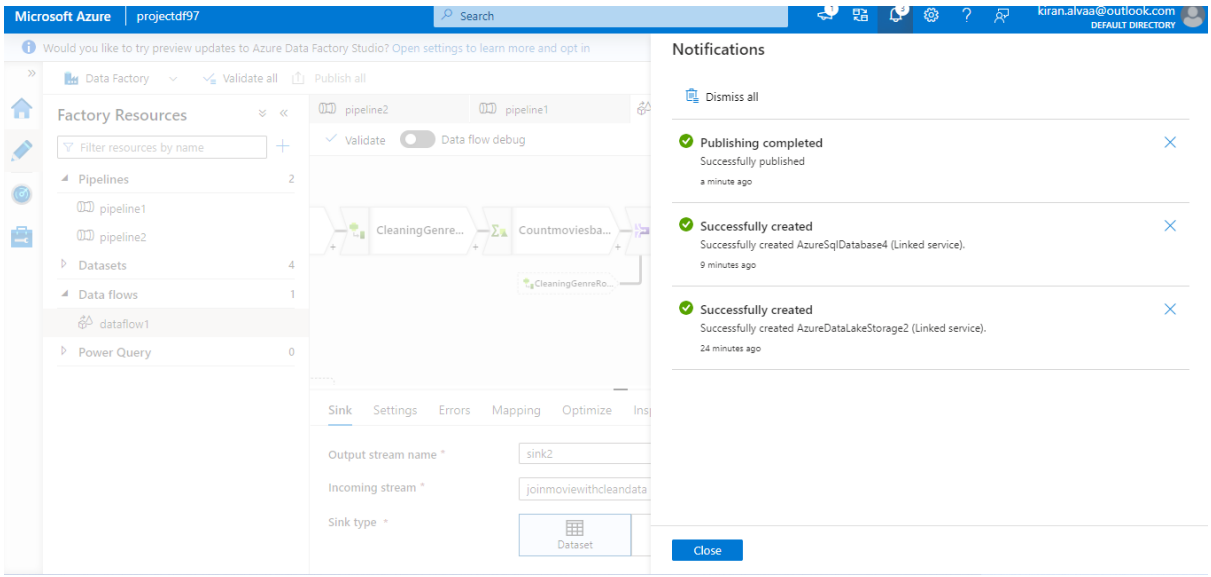
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

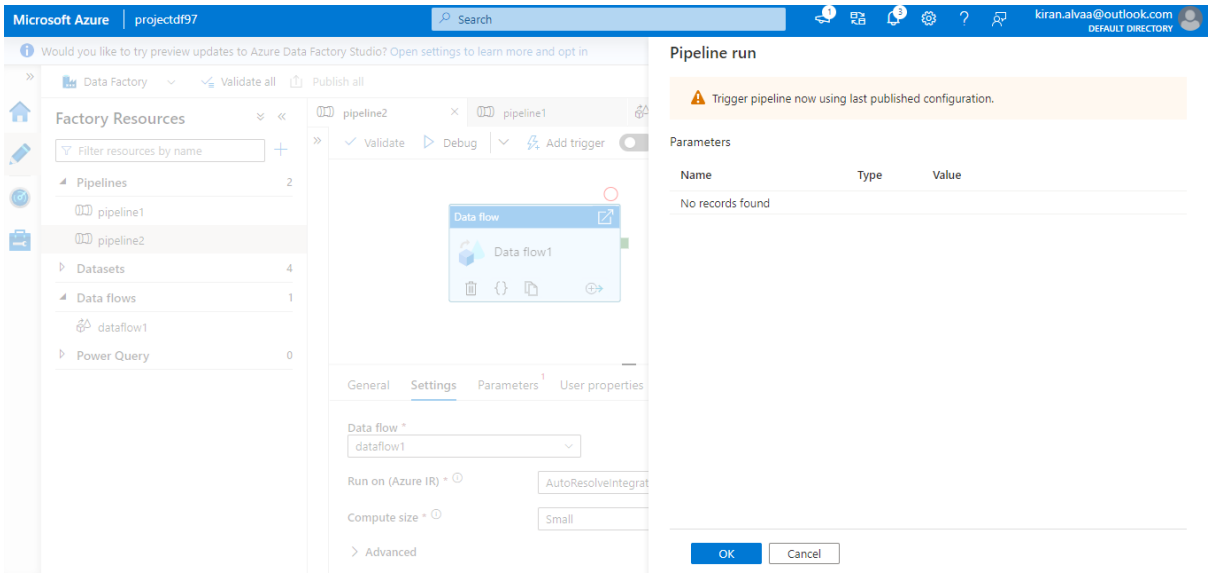




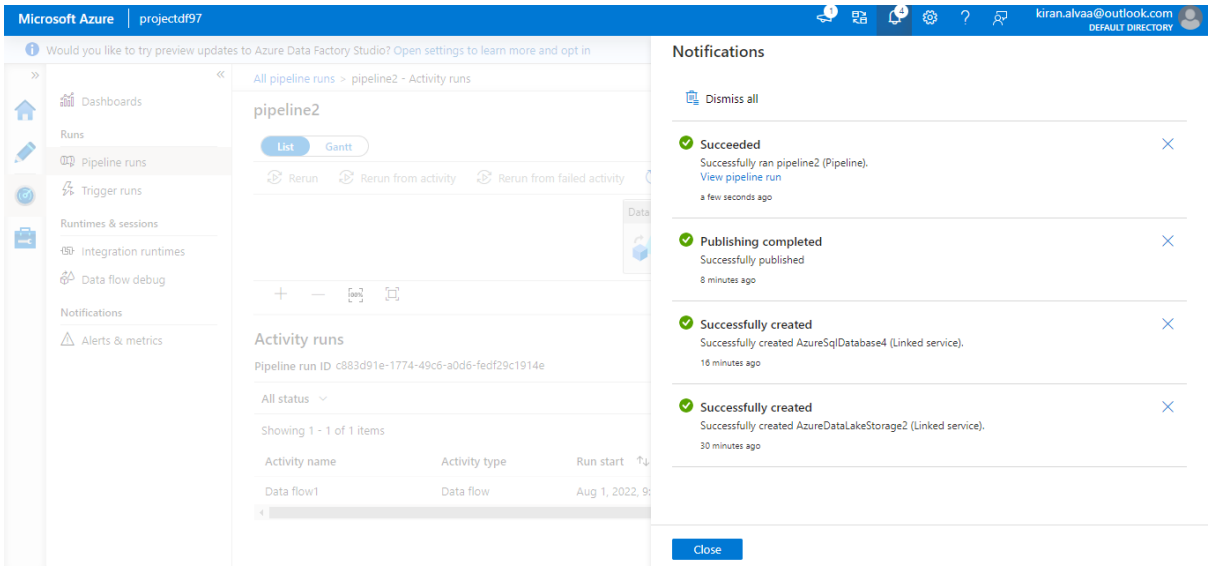
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

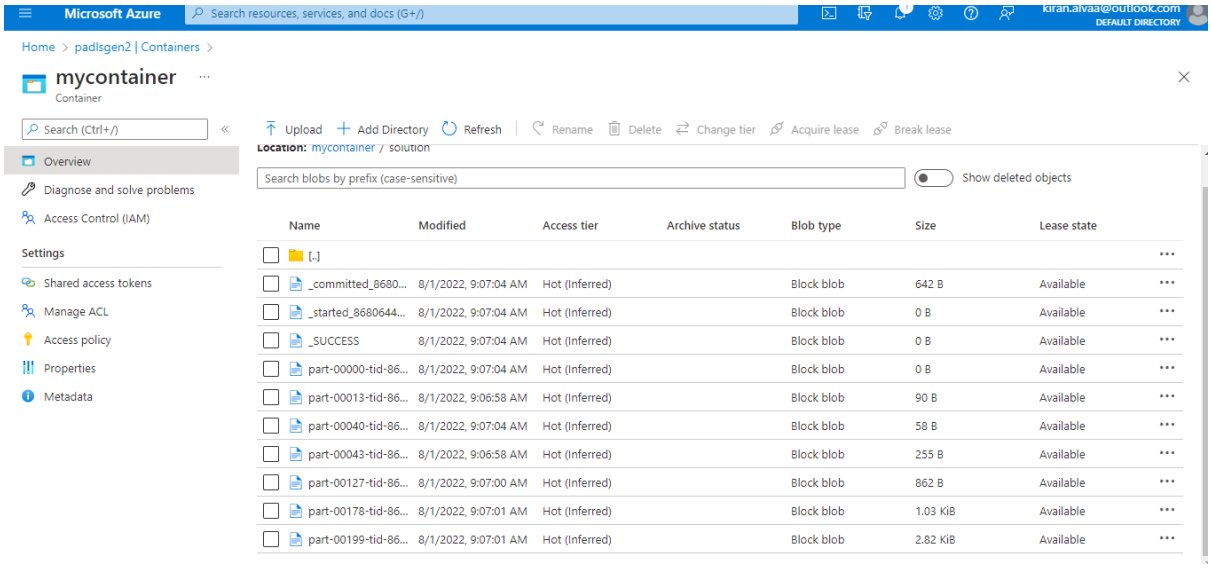






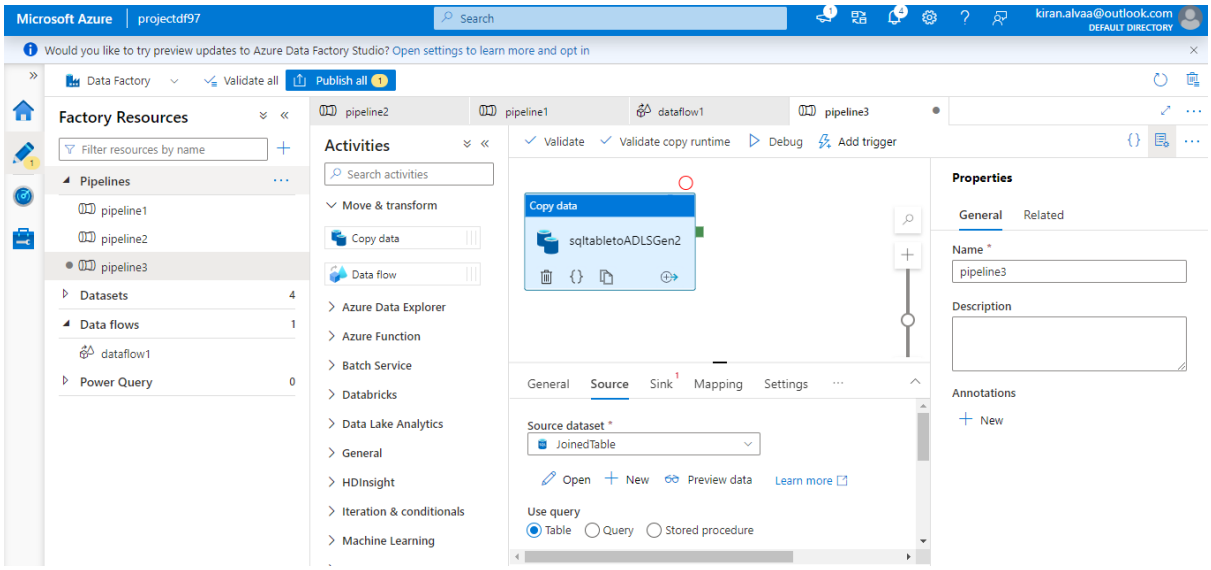
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

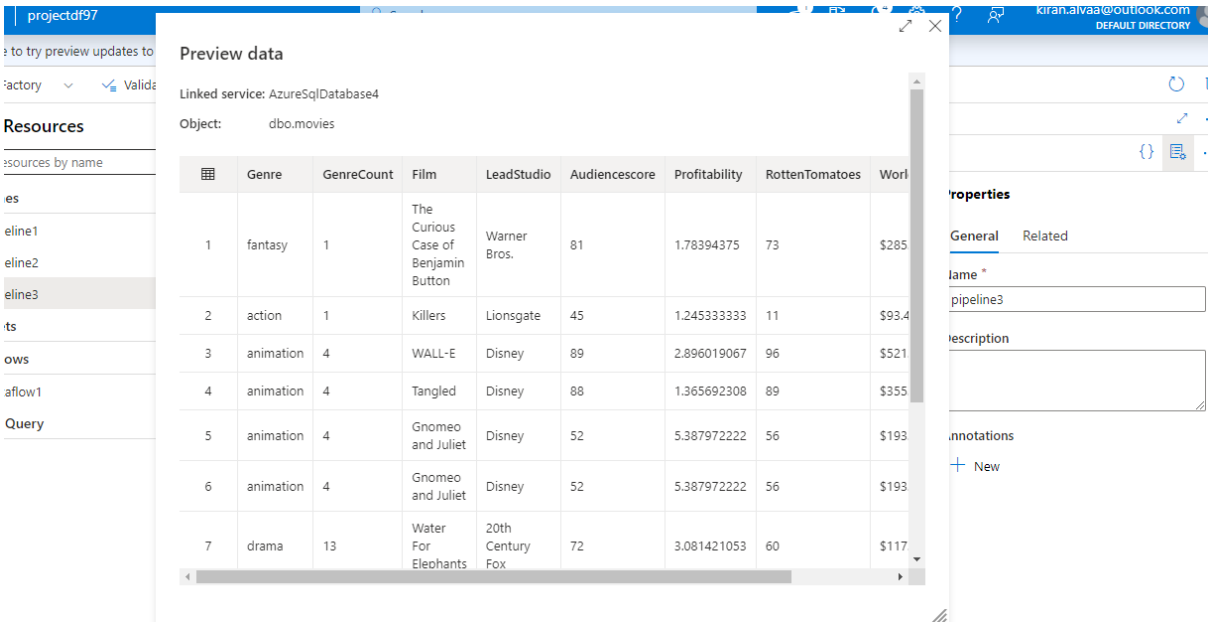


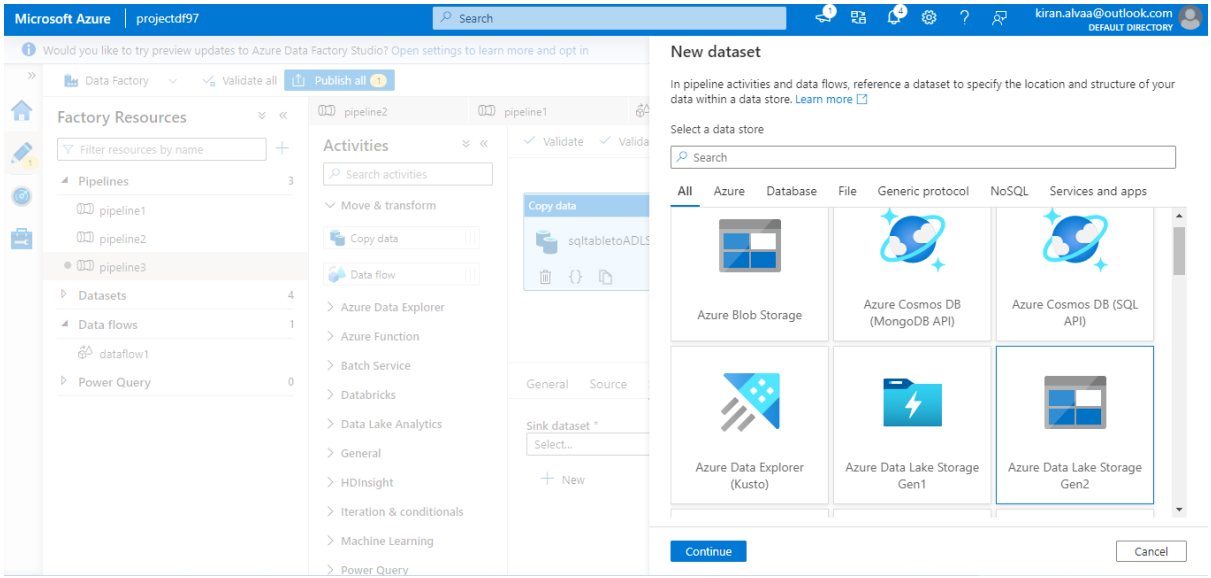


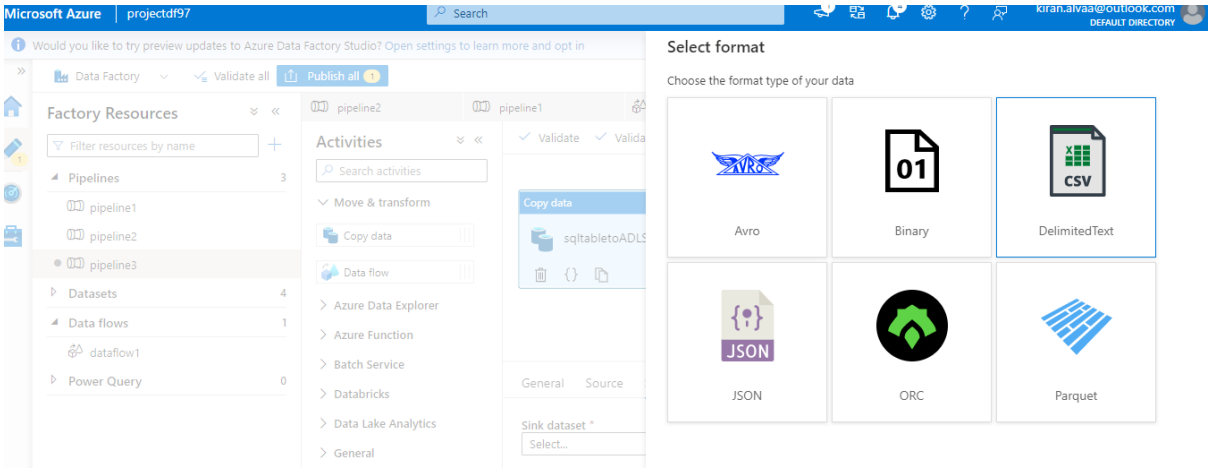
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

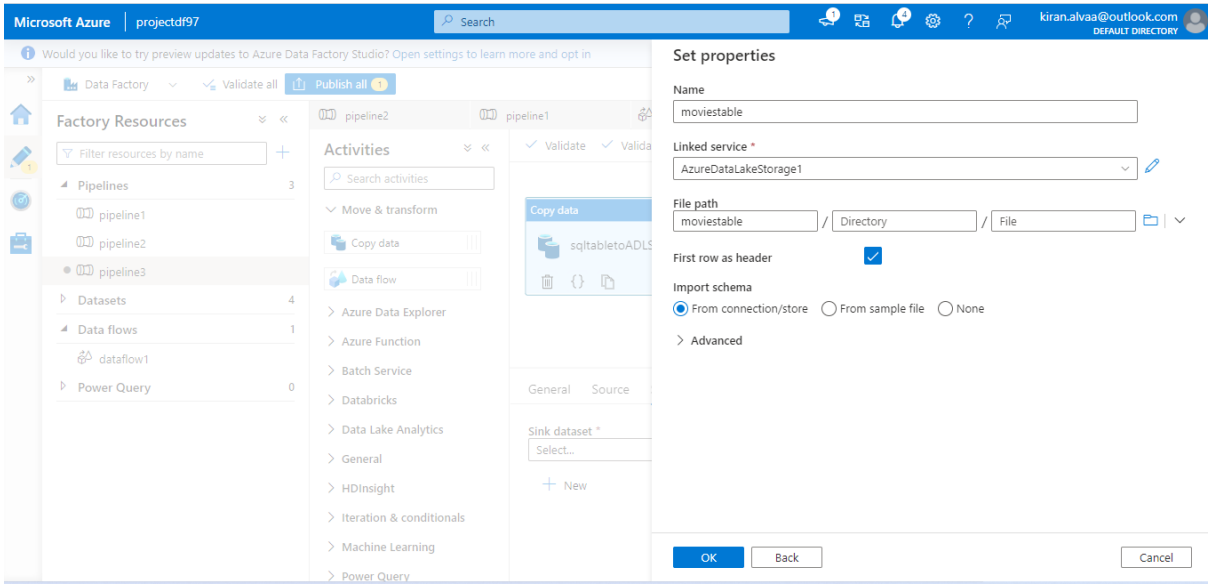
1. Get the clean data from Azure SQL DB. Create an activity that can copy the data from SQLDB to ADLS Gen2.





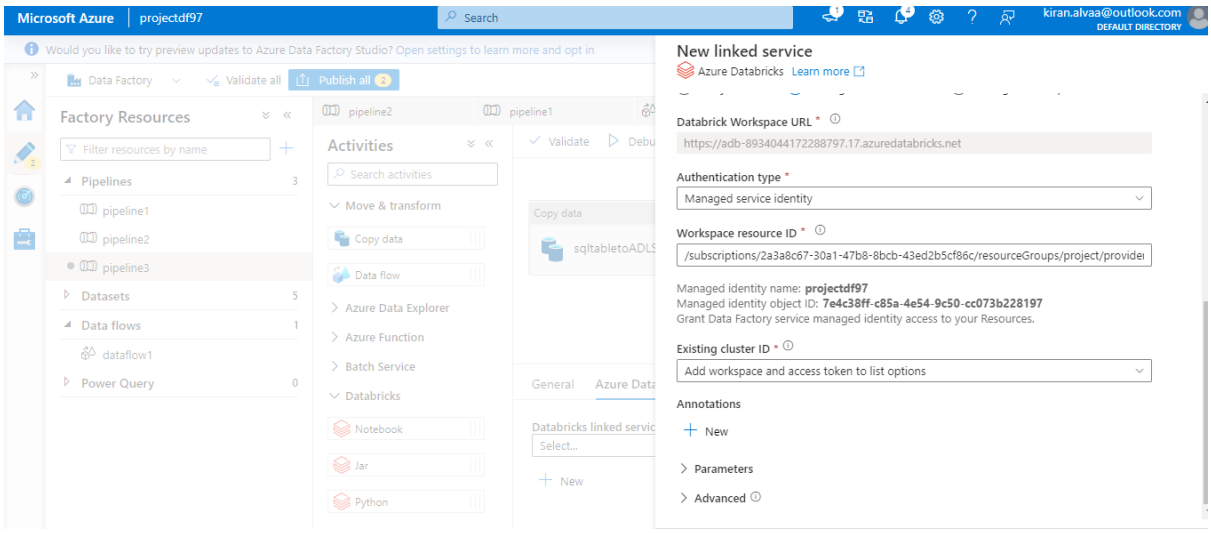


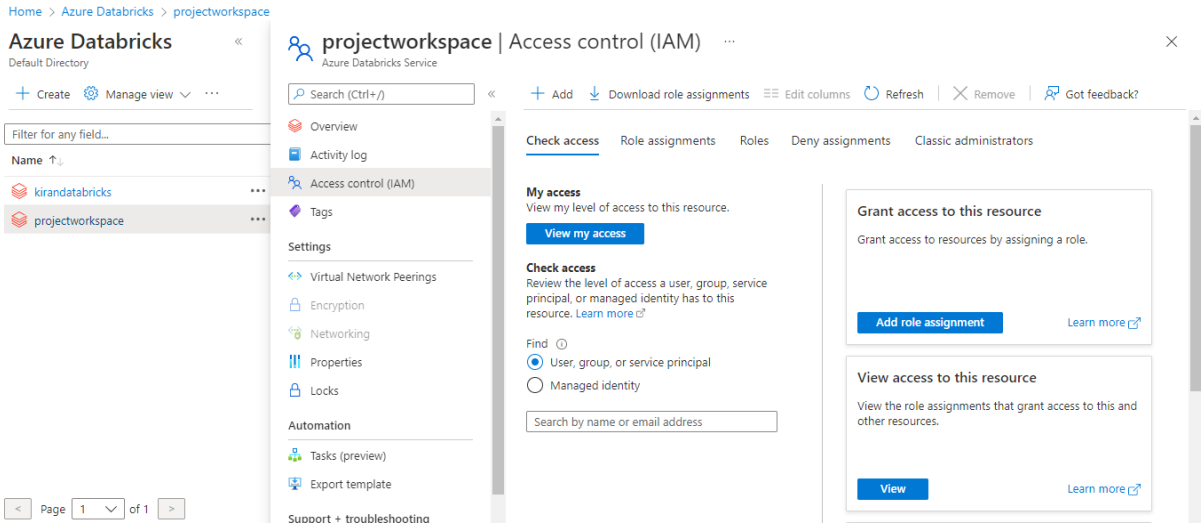


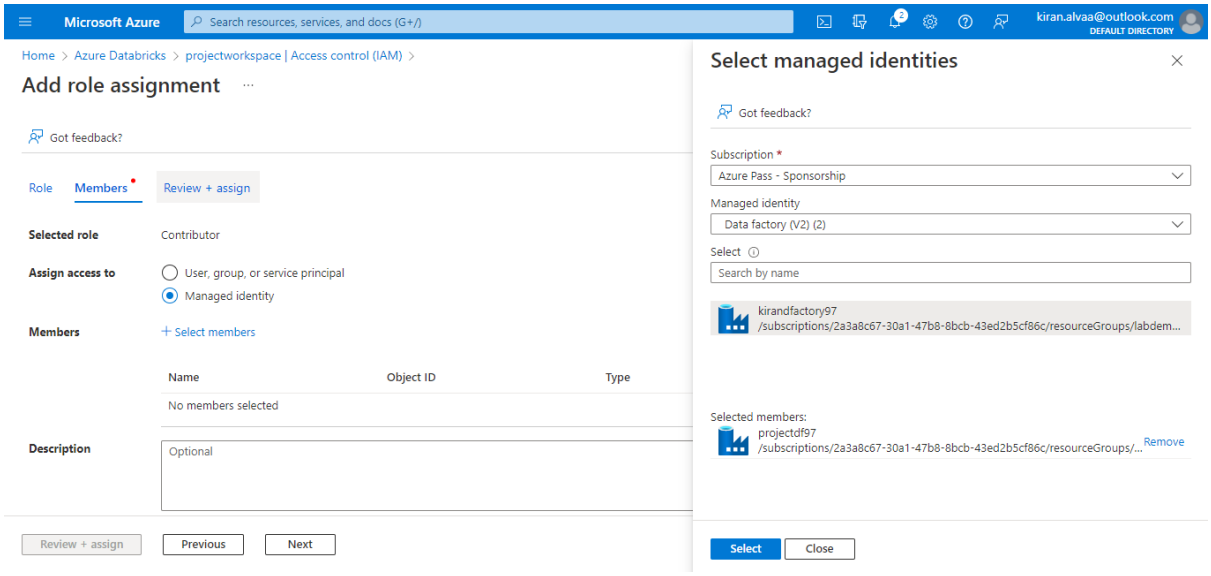


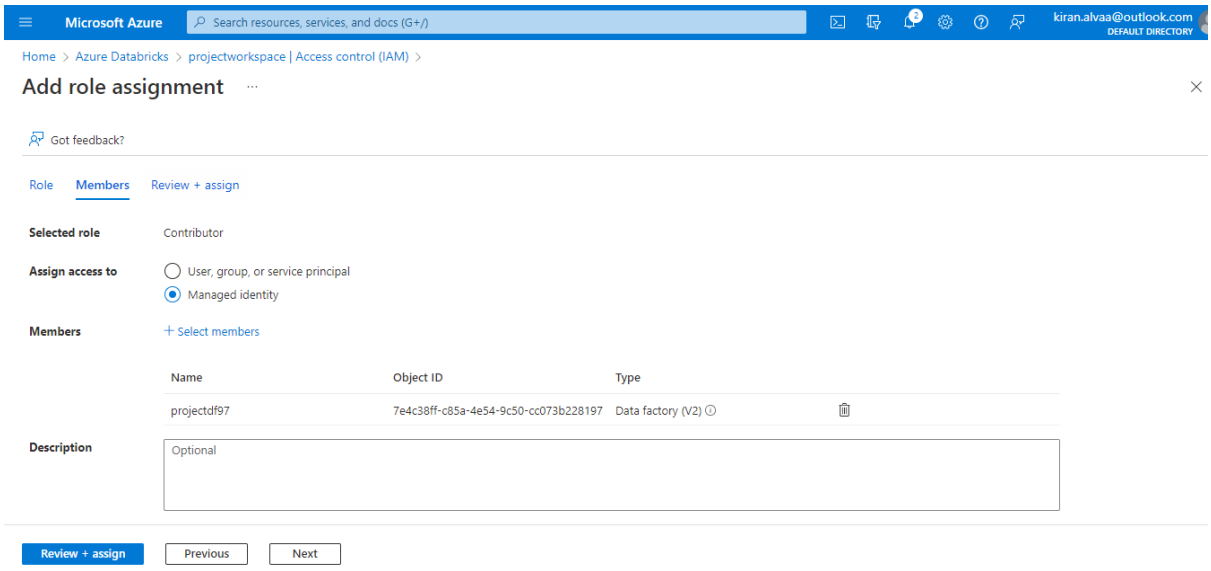
2. Create an activity that can use Azure Databricks to read the data from the ADLS Gen2 and perform rank operation on the Genre column. Ensure this activity gets activated only after the data is stored in ADLS from SQL DB. The result of Databricks must be stored in the ADLS.

Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

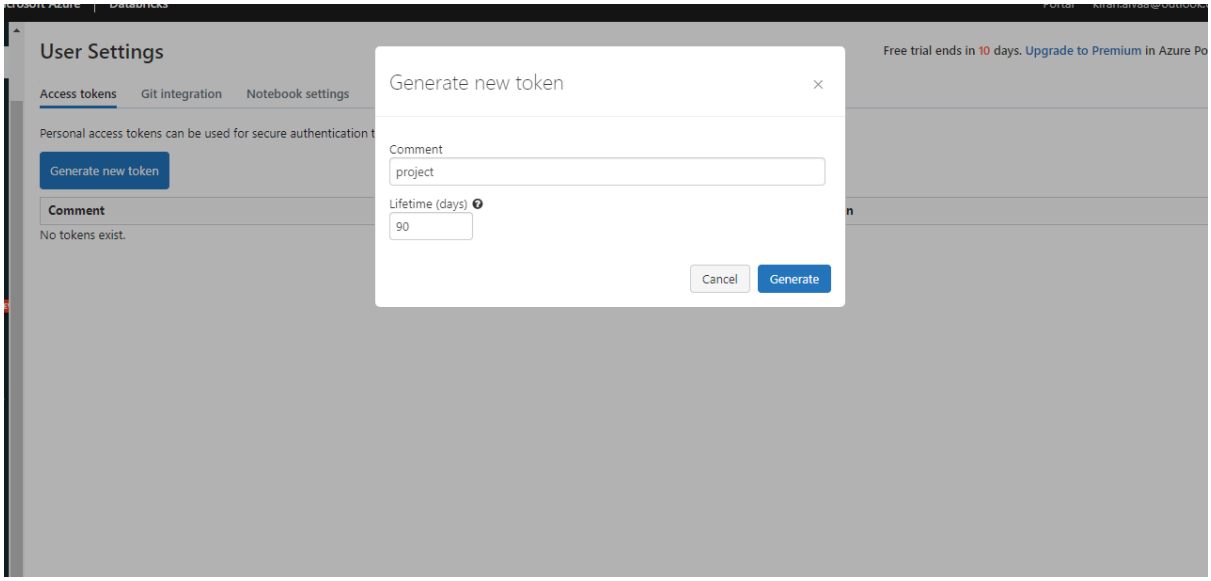


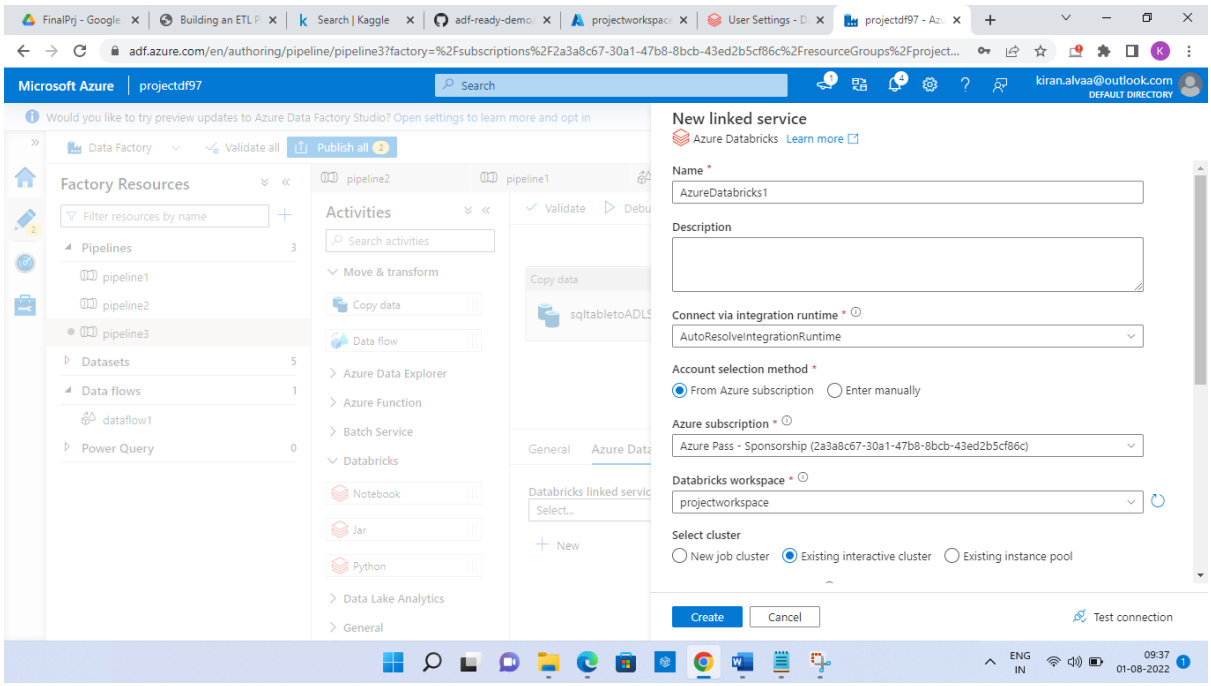


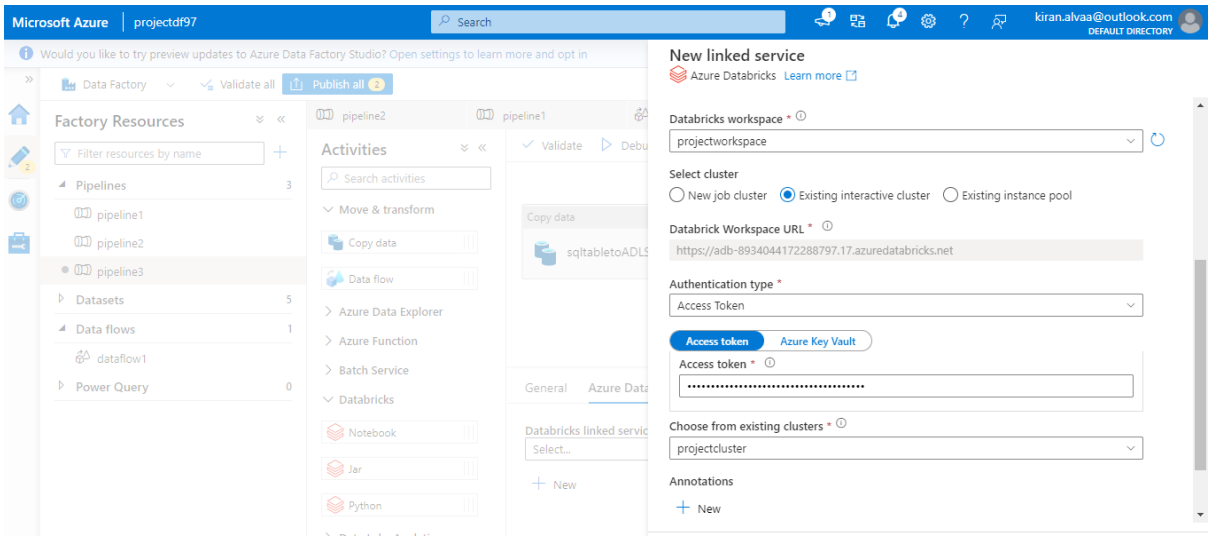

Microsoft Azure
Search resources, services, and docs (G+/)
kiran.alvaa@outlook.com
DEFAULT DIRECTORY
Home > Azure Databricks > projectworkspace | Access control (IAM) >
Add role assignment
Got feedback?
Role
Members
Review + assign
Selected role
Contributor
Assign access to
User, group, or service principal
Managed identity
Members
+ Select members
Name
Object ID
Type
No members selected
Description
Optional
Review + assign
Previous
Next
Select managed identities
Got feedback?
Subscription *
Azure Pass - Sponsorship
Managed identity
Data factory (V2) (2)
Select
Search by name
kirandfactory97
/subscriptions/2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c/resourceGroups/labdem...
Selected members:
projectdf97
/subscriptions/2a3a8c67-30a1-47b8-8bcb-43ed2b5cf86c/resourceGroups/...
Remove
Select
Close


Microsoft Azure
Search resources, services, and docs (G+/)
kiran.alvaa@outlook.com
DEFAULT DIRECTORY
Home > Azure Databricks > projectworkspace | Access control (IAM) >
Add role assignment
Got feedback?
Role
Members
Review + assign
Selected role
Contributor
Assign access to
User, group, or service principal
Managed identity
Members
+ Select members
Name
Object ID
Type
projectdf97
7e4c38ff-c85a-4e54-9c50-cc073b228197
Data factory (V2)
Description
Optional
Review + assign
Previous
Next

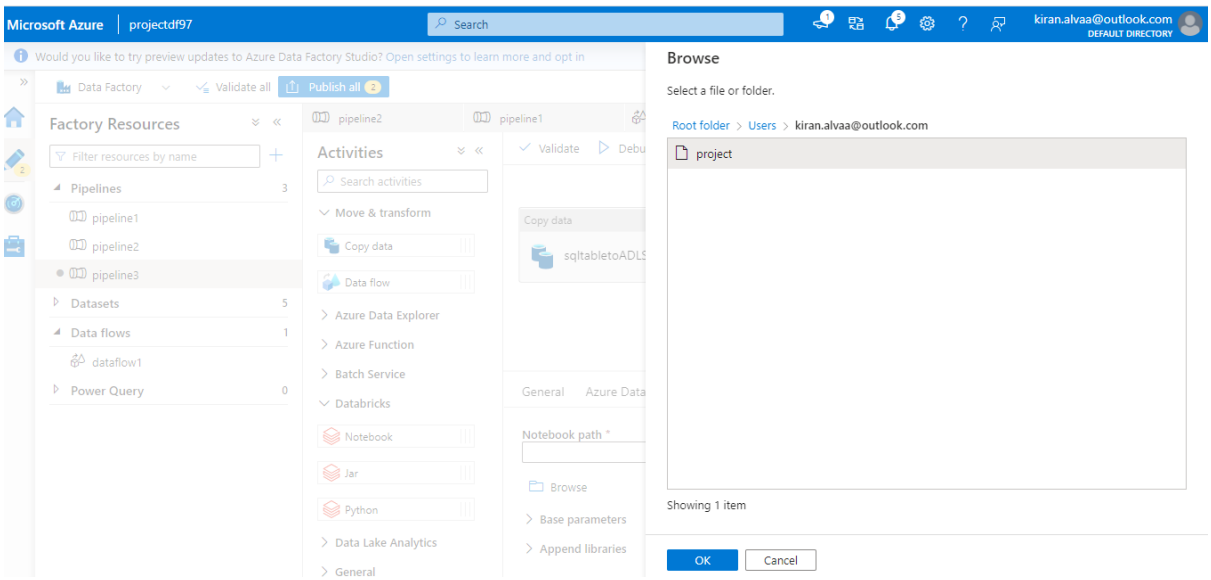
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

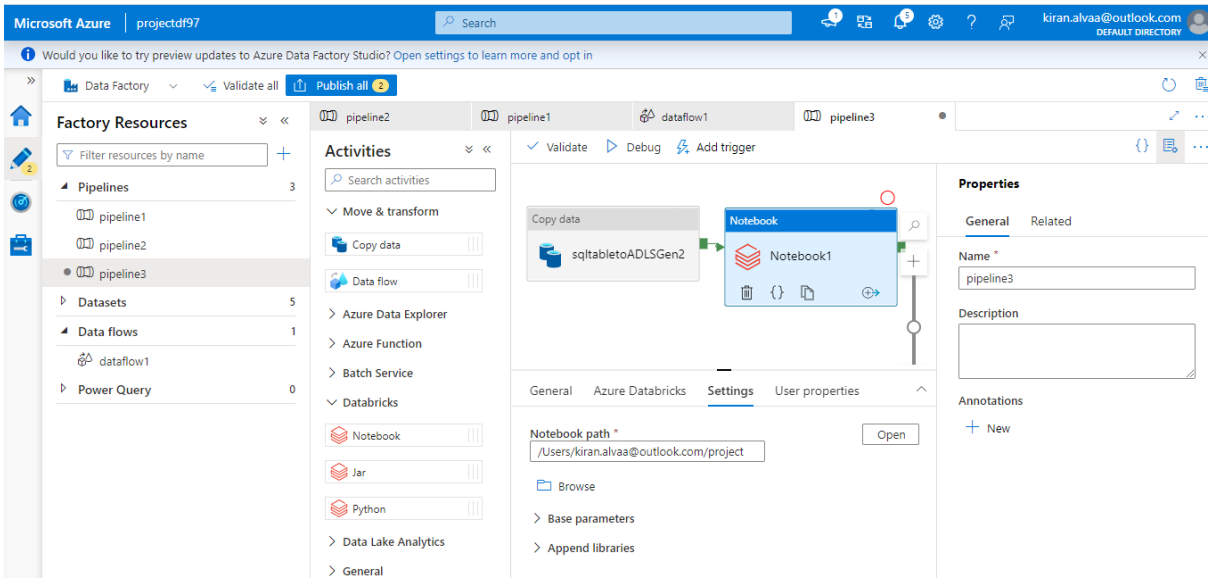


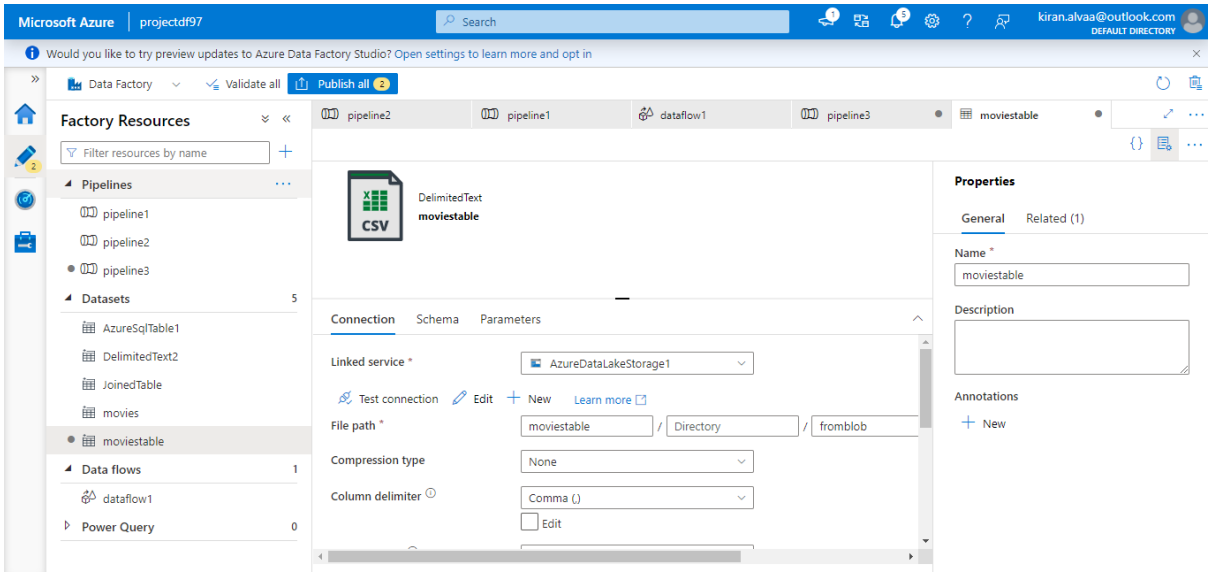


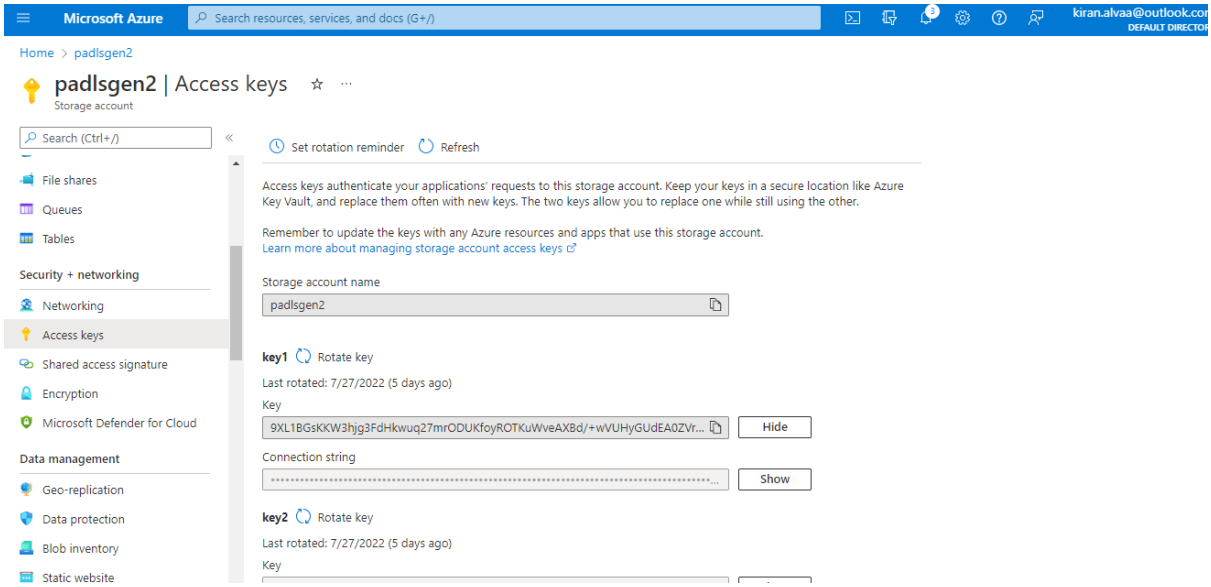


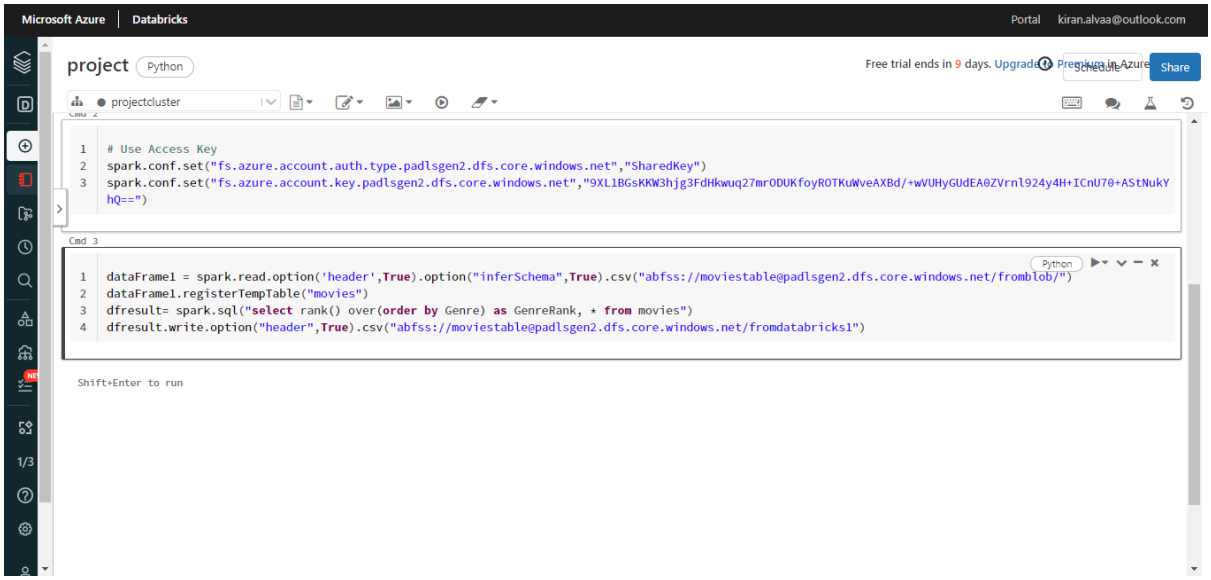
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services











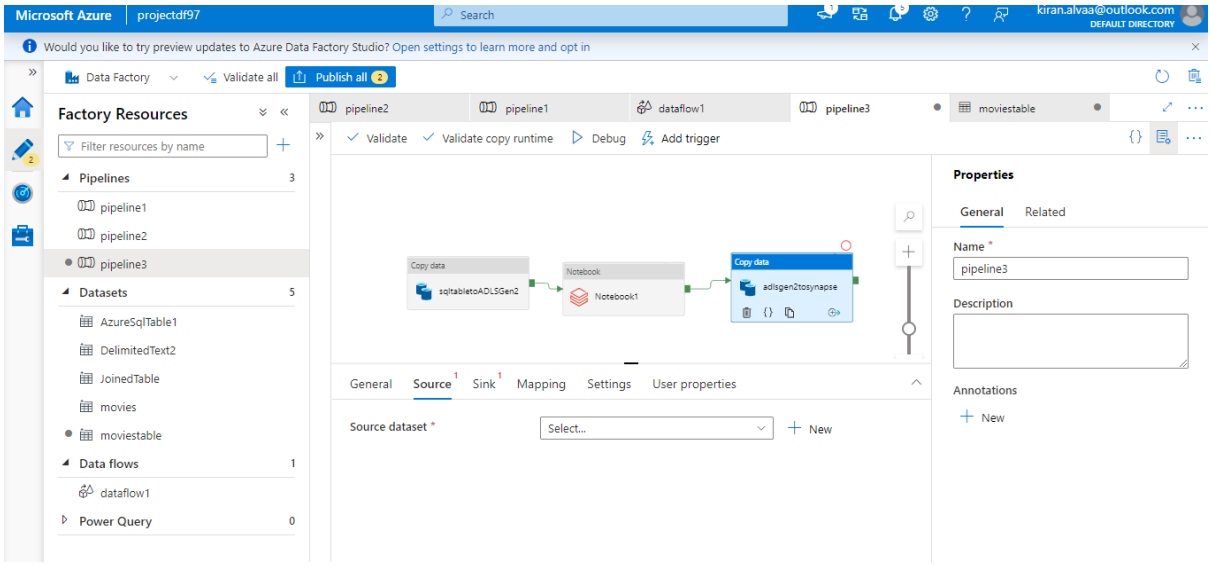
dataFrame1 = spark.read.option('header',True).option("inferSchema",True).csv("abfss://moviestable@padlsgen2.dfs.core.windows.net/fromblob/")

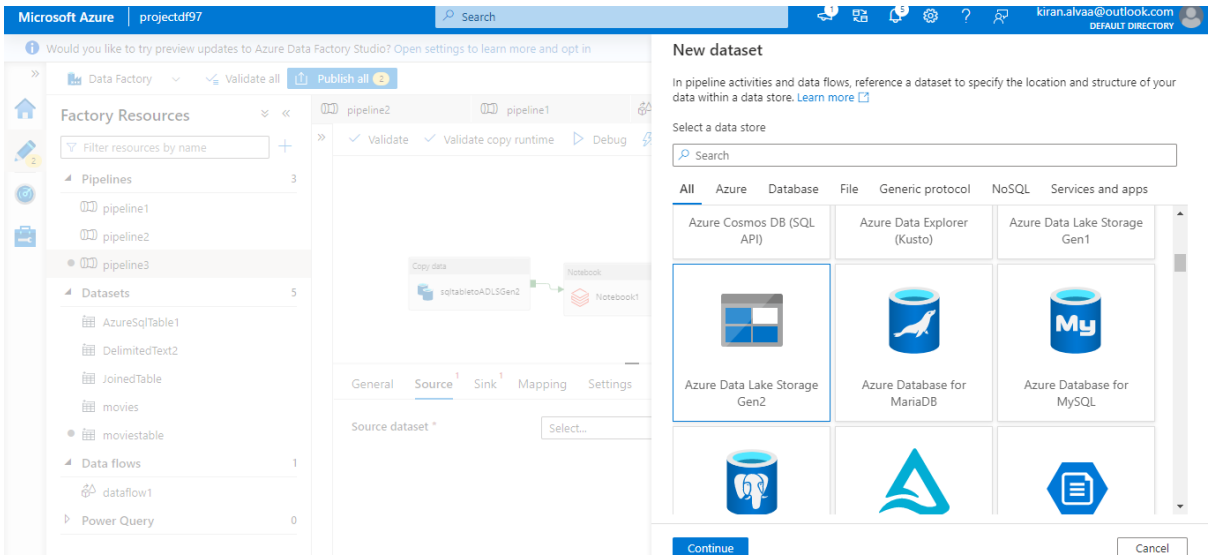
dataFrame1.registerTempTable("movies")

dfresult= spark.sql("select rank() over(order by Genre) as GenreRank, * from movies")

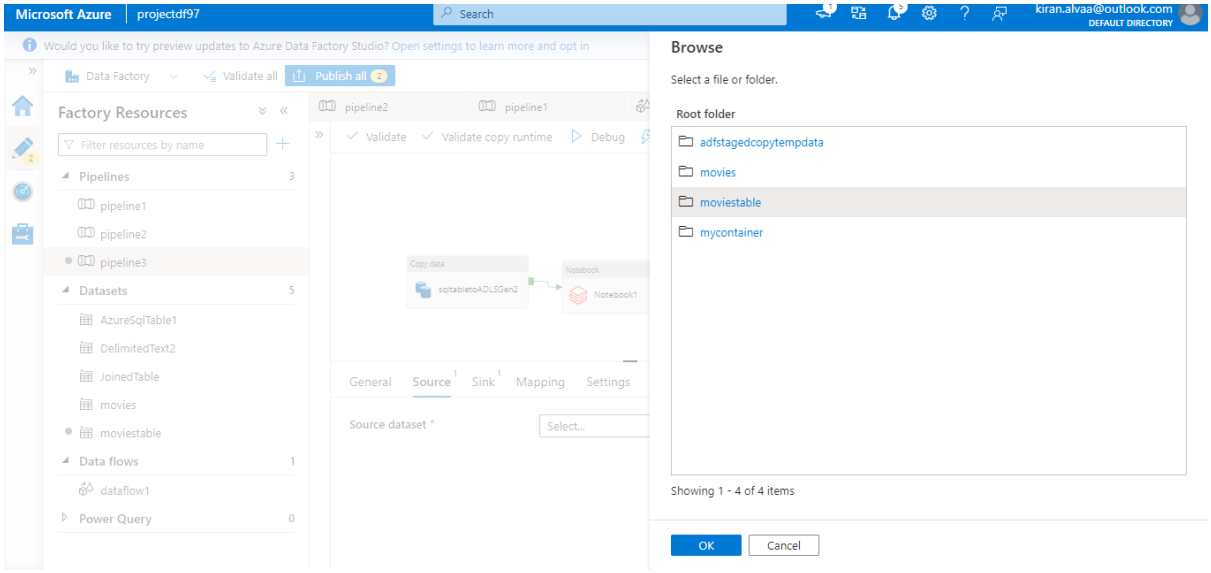
dfresult.write.option("header",True).csv("abfss://moviestable@padlsgen2.dfs.core.windows.net/fromdatabricks1")

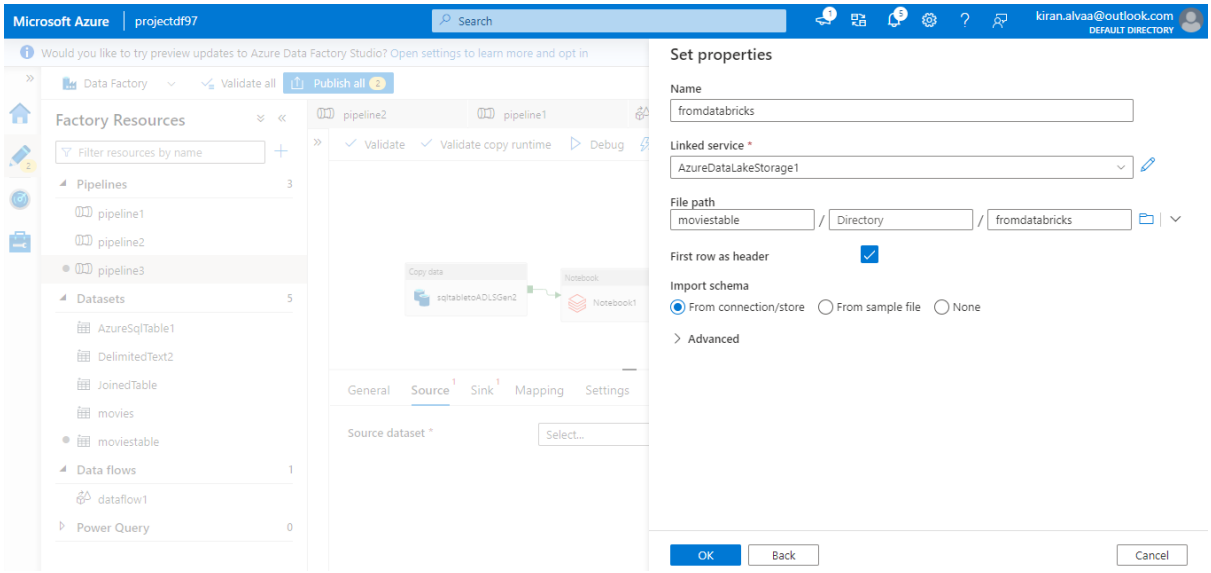
3. Create a final activity that will read the output of previous activity in ADLS and store the same in Synapse.



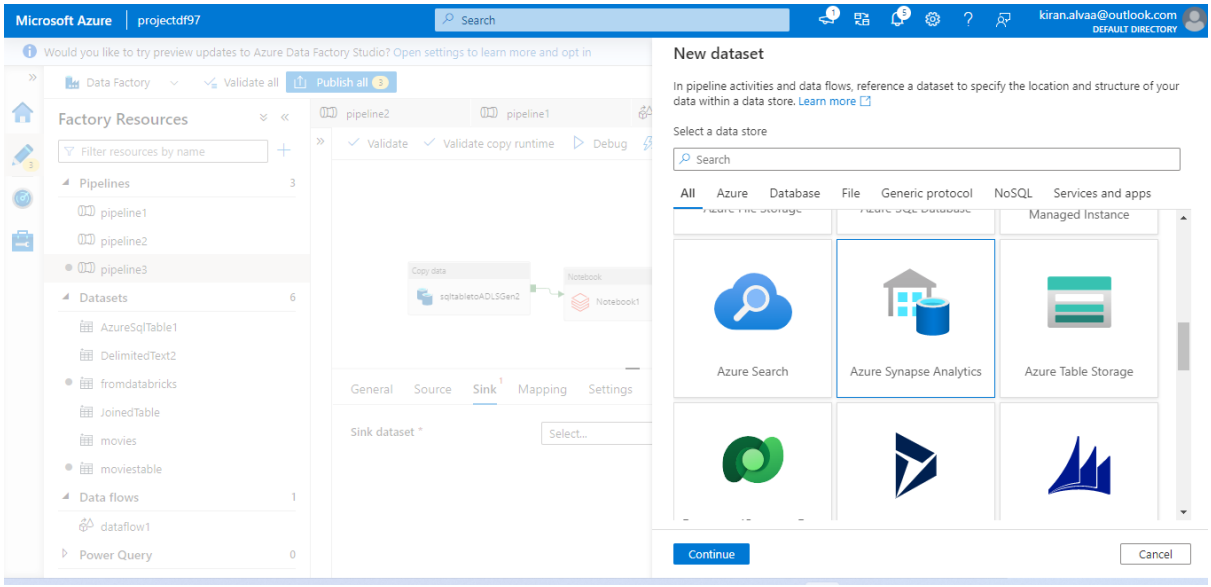


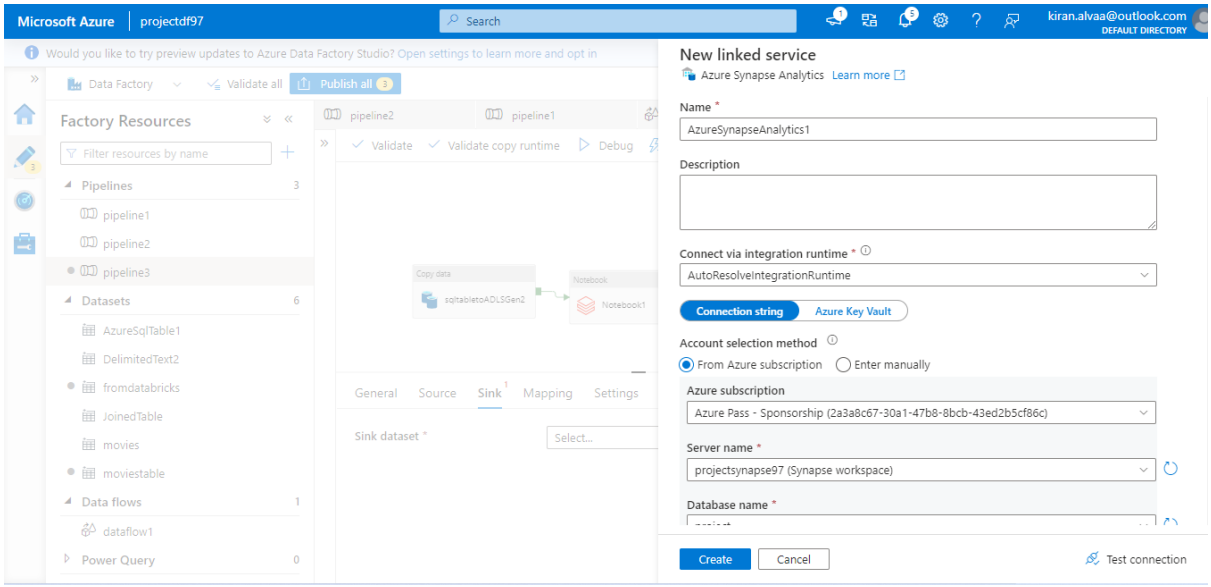
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

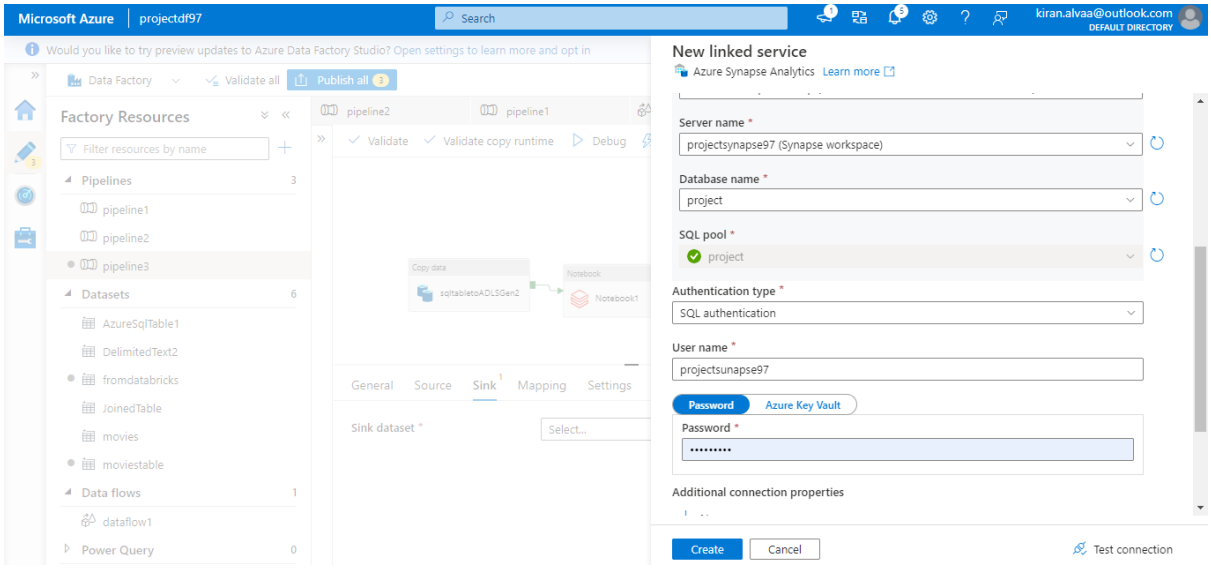


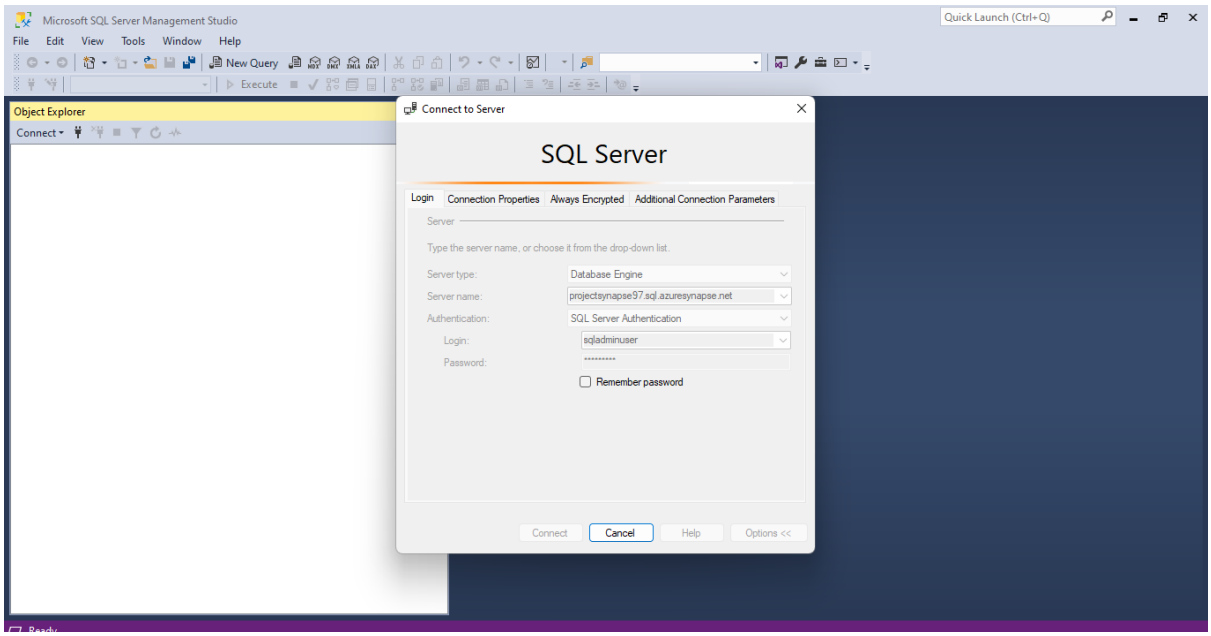


Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services



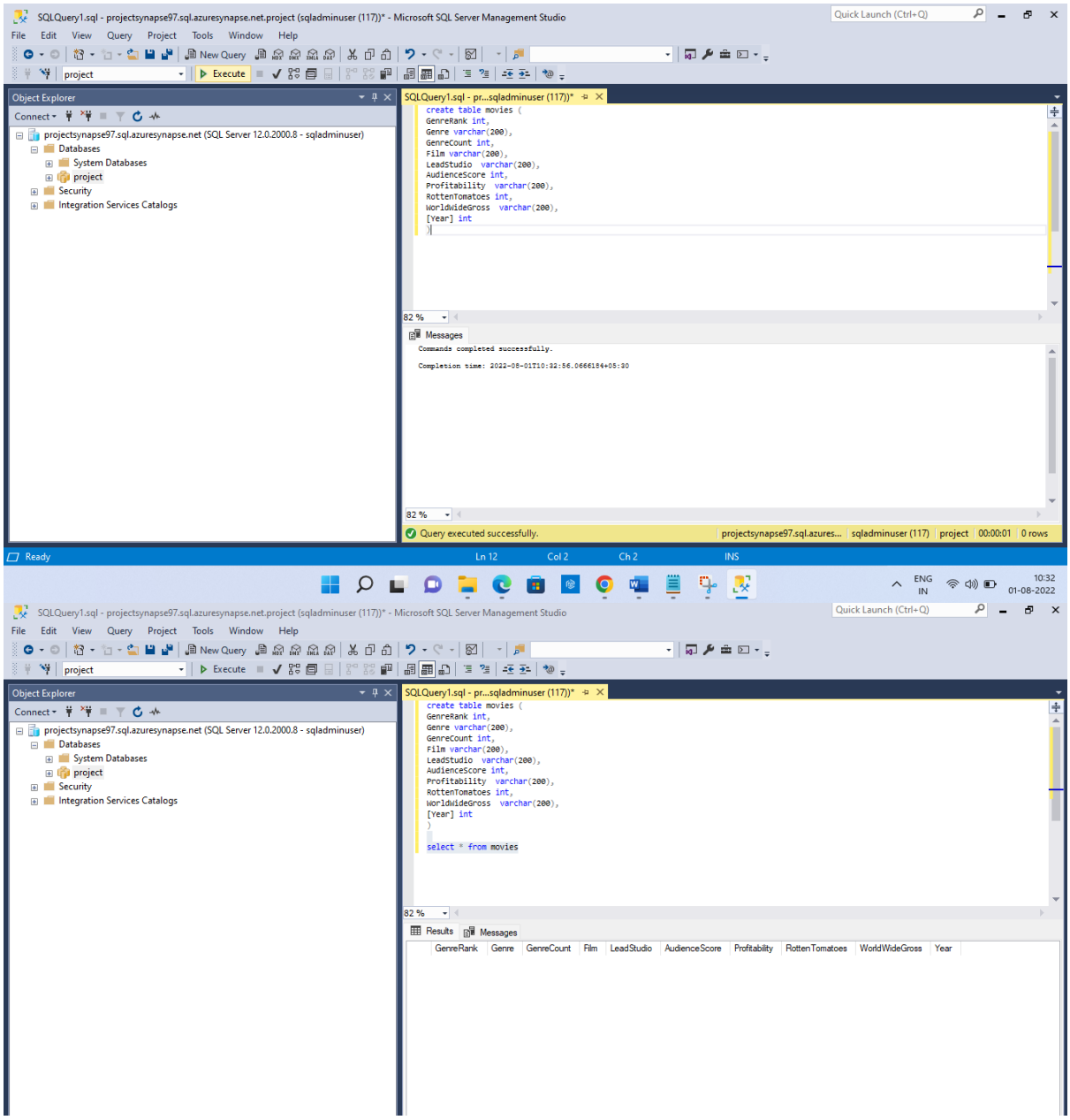


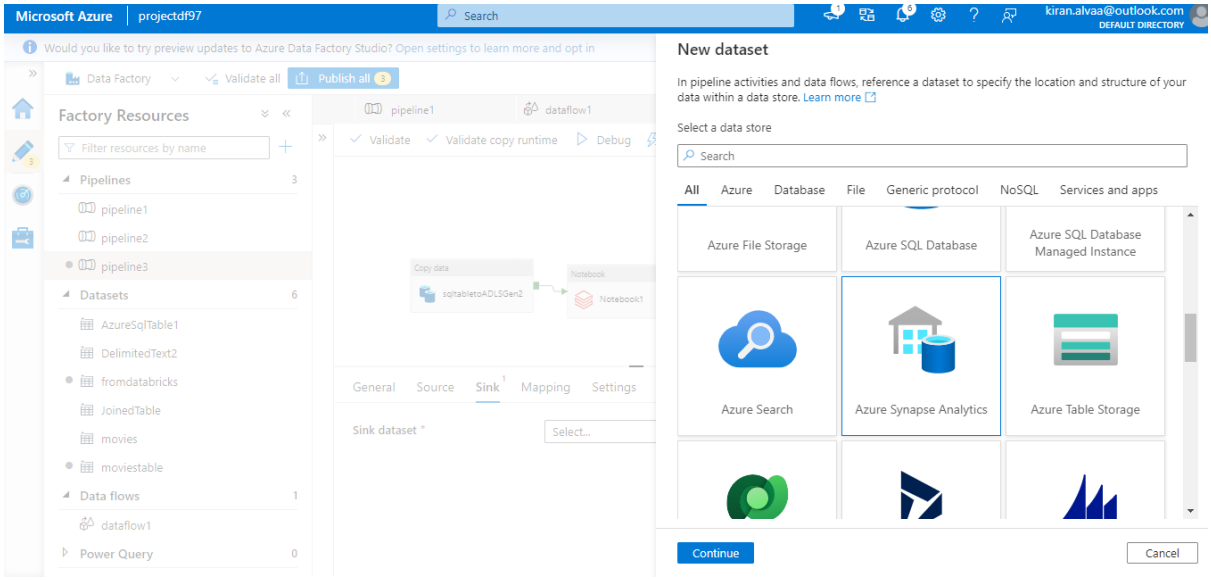


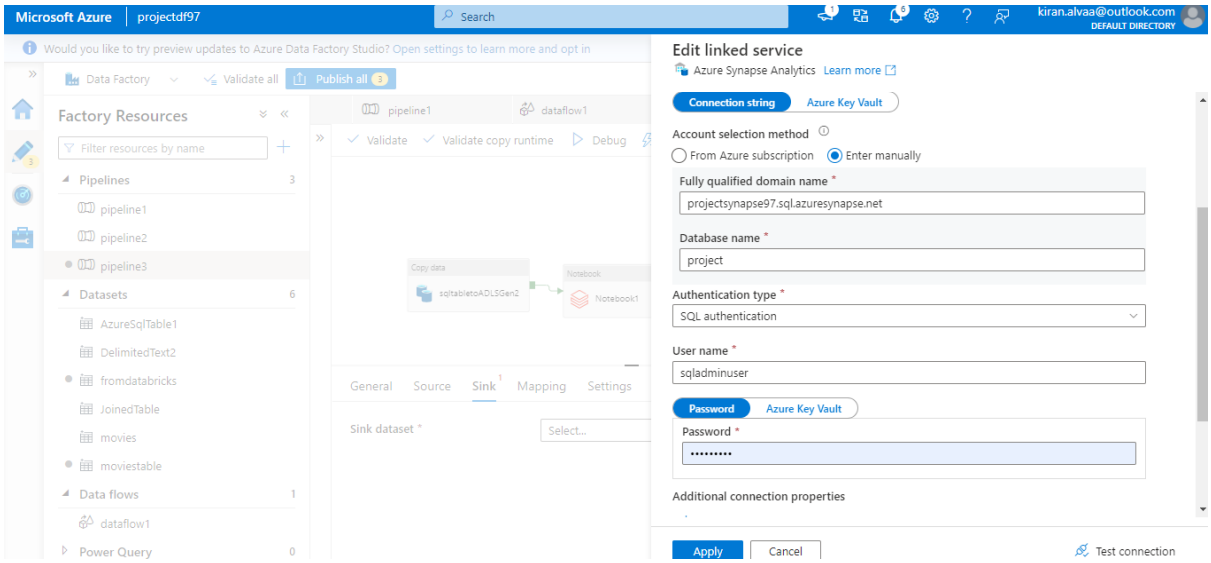


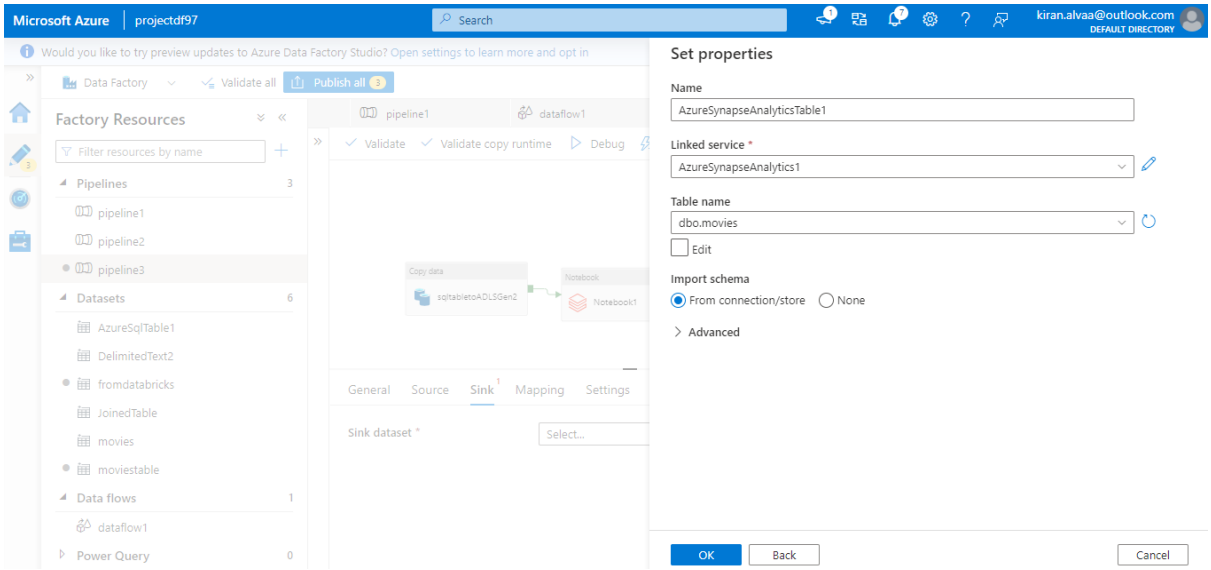

```
create table movies (
GenreRank int,
Genre varchar(200),
GenreCount int,
Film varchar(200),
LeadStudio  varchar(200),
AudienceScore int,
Profitability  varchar(200),
RottenTomatoes int,
WorldWideGross  varchar(200),
```

```
[Year] int
                                                                    )
```

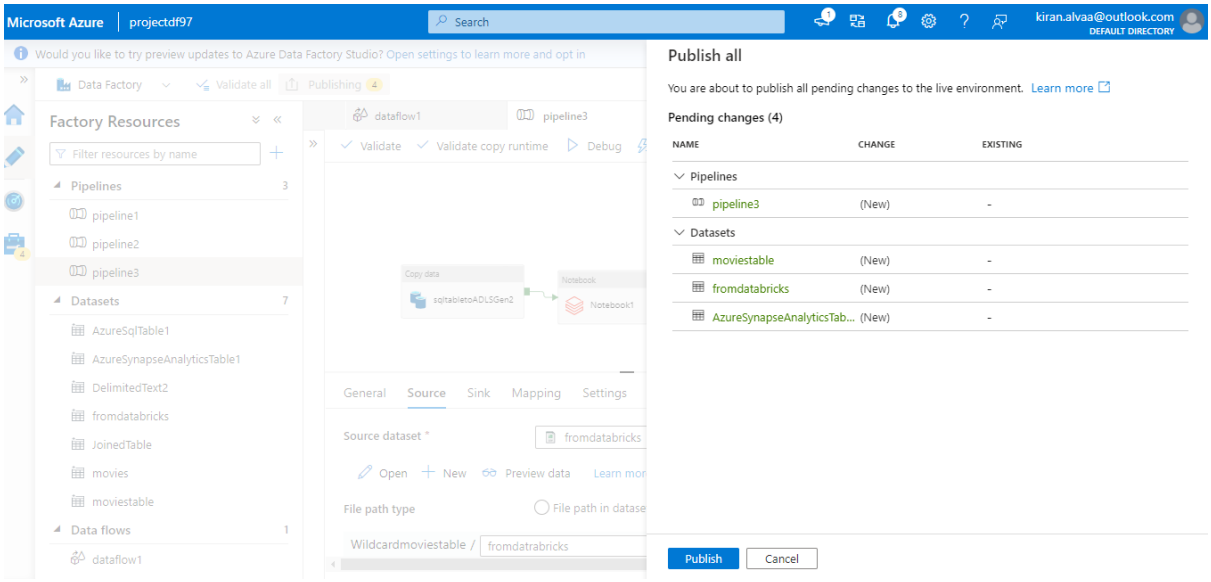

SQLQuery1.sql - projectsynapse97.sql.azuresynapse.net.project (sqladminuser (117))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Tools Window Help
New Query
Execute
project
Object Explorer
Connect
projectsynapse97.sql.azuresynapse.net (SQL Server 12.0.2000.8 - sqladminuser)
Databases
System Databases
project
Security
Integration Services Catalogs
SQLQuery1.sql - pr...sqladminuser (117))*
create table movies (
GenreRank int,
Genre varchar(200),
GenreCount int,
Film varchar(200),
LeadStudio varchar(200),
AudienceScore int,
Profitability varchar(200),
RottenTomatoes int,
WorldWideGross varchar(200),
[Year] int
82 %
Messages
Commands completed successfully.
Completion time: 2022-08-01T10:32:56.0666184+05:30
Query executed successfully.
projectsynapse97.sql.azures...
sqladminuser (117)
project
00:00:01
0 rows
Ready
Ln 12
Col 2
Ch 2
INS
ENG
IN
10:32
01-08-2022
select * from movies
Results
Messages
GenreRank
Genre
GenreCount
Film
LeadStudio
AudienceScore
Profitability
RottenTomatoes
WorldWideGross
Year

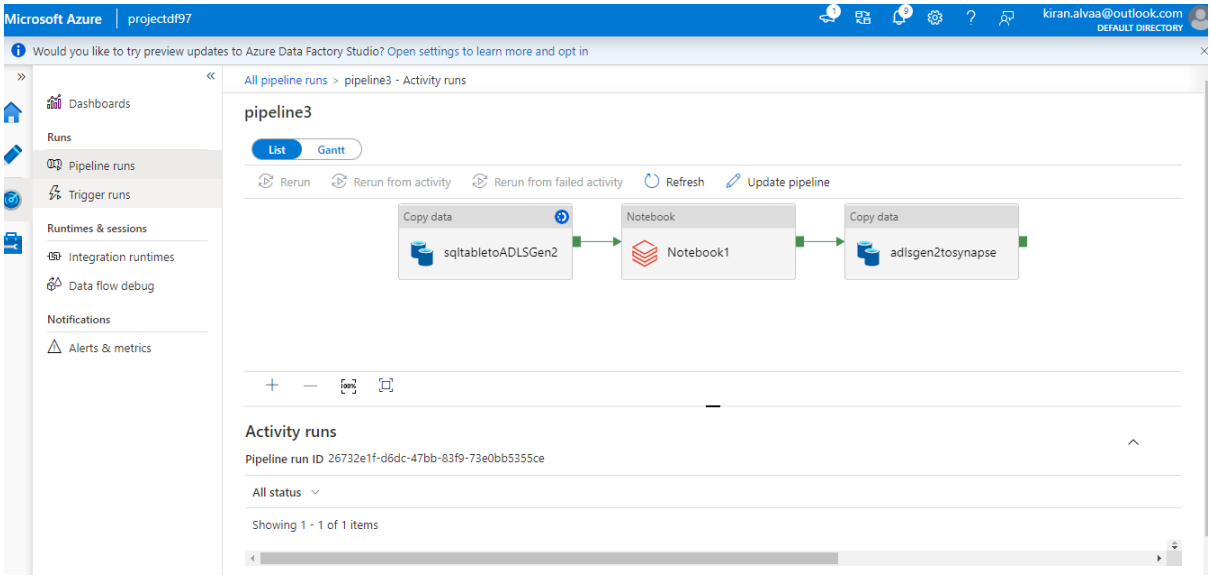

Microsoft Azure
projectdf97
Search
kiran.alvaa@outlook.com
DEFAULT DIRECTORY
Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in
Data Factory
Validate all
Publish all 3
Factory Resources
Filter resources by name
Pipelines 3
pipeline1
pipeline2
pipeline3
Datasets 6
AzureSqlTable1
DelimitedText2
fromdatabricks
JoinedTable
movies
moviestable
Data flows 1
dataflow1
Power Query 0
pipeline1
dataflow1
Validate
Validate copy runtime
Debug
General
Source
Sink 1
Mapping
Settings
Sink dataset *
Select...
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 File Storage
Azure SQL Database
Azure SQL Database Managed Instance
Azure Search
Azure Synapse Analytics
Azure Table Storage
Continue
Cancel

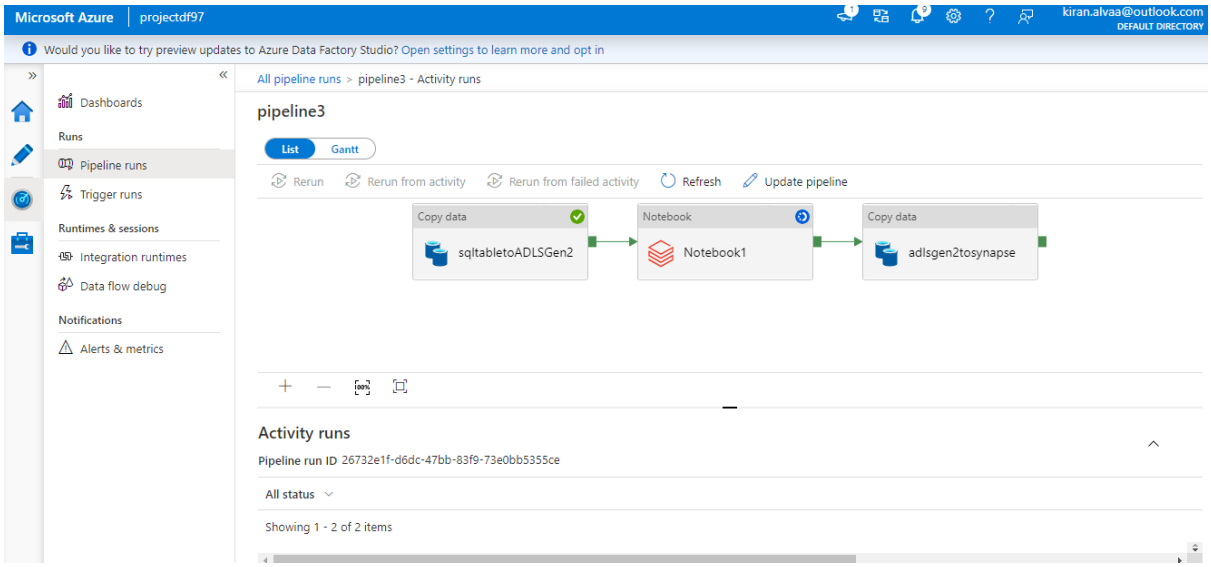

Microsoft Azure
projectdf97
Search
kiran.alvaa@outlook.com
DEFAULT DIRECTORY
Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in
Data Factory
Validate all
Publish all 3
Factory Resources
Filter resources by name
Pipelines 3
pipeline1
pipeline2
pipeline3
Datasets 6
AzureSqlTable1
DelimitedText2
fromdatabricks
JoinedTable
movies
moviestable
Data flows 1
dataflow1
Power Query 0
pipeline1
dataflow1
Validate
Validate copy runtime
Debug
General
Source
Sink 1
Mapping
Settings
Sink dataset *
Select...
Edit linked service
Azure Synapse Analytics Learn more
Connection string
Azure Key Vault
Account selection method
From Azure subscription
Enter manually
Fully qualified domain name *
projectsynapse97.sql.azuresynapse.net
Database name *
project
Authentication type *
SQL authentication
User name *
sqladminuser
Password
Azure Key Vault
Password *
Additional connection properties
Apply
Cancel
Test connection


Microsoft Azure
projectdf97
Search
kiran.alvaa@outlook.com
DEFAULT DIRECTORY
Would you like to try preview updates to Azure Data Factory Studio? Open settings to learn more and opt in
Data Factory
Validate all
Publish all 3
Factory Resources
Filter resources by name
Pipelines 3
pipeline1
pipeline2
pipeline3
Datasets 6
AzureSqlTable1
DelimitedText2
fromdatabricks
JoinedTable
movies
moviestable
Data flows 1
dataflow1
Power Query 0
pipeline1
dataflow1
Validate
Validate copy runtime
Debug
General
Source
Sink 1
Mapping
Settings
Sink dataset *
Select...
Set properties
Name
AzureSynapseAnalyticsTable1
Linked service *
AzureSynapseAnalytics1
Table name
dbo.movies
Edit
Import schema
From connection/store
None
Advanced
OK
Back
Cancel

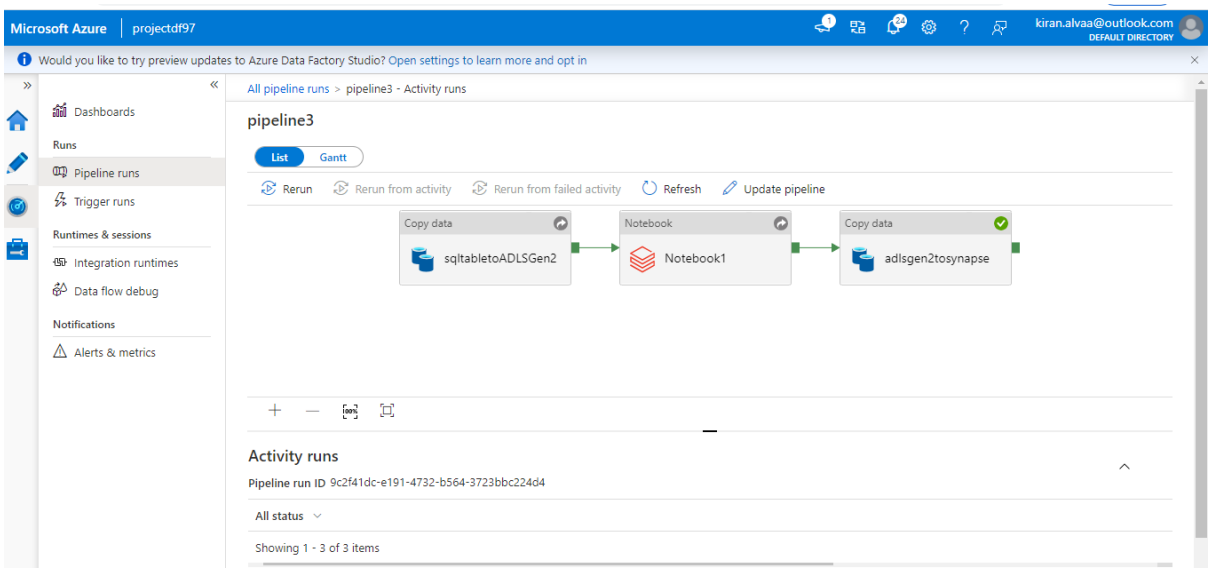
Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services

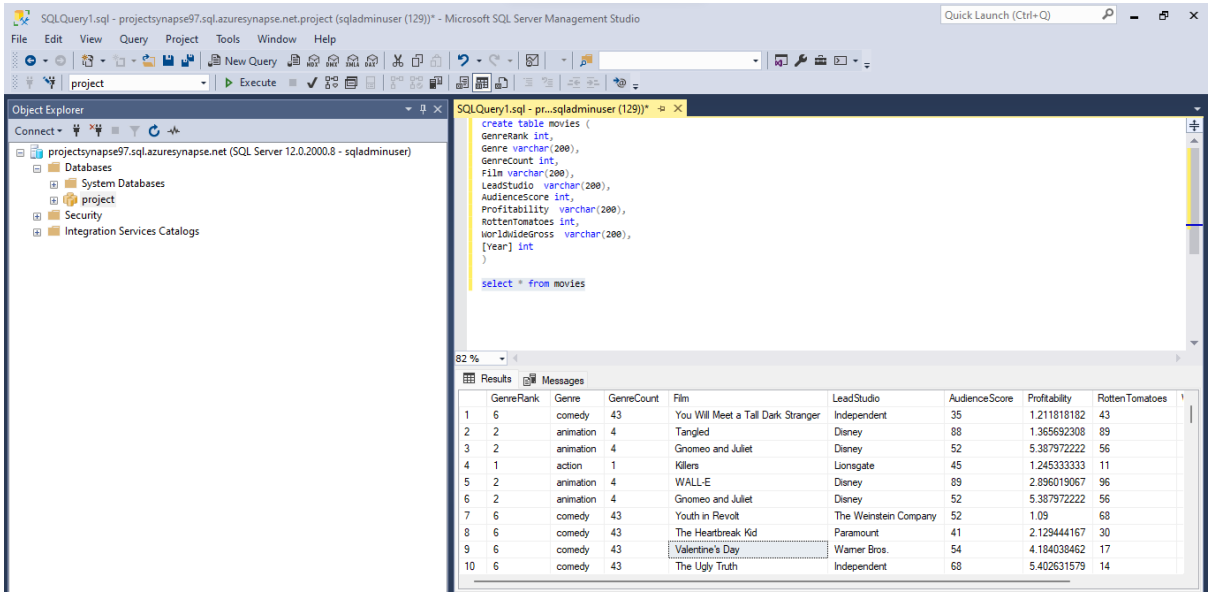






Kiran Shridhar Alva -Building an ETL Pipeline using Azure Data Services





	GenreRank	Genre	GenreCount	Film	LeadStudio	AudienceScore	Profitability	RottenTomatoes
1	6	comedy	43	You Will Meet a Tall Dark Stranger	Independent	35	1.211818182	43
2	2	animation	4	Tangled	Disney	88	1.365692308	89
3	2	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56
4	1	action	1	Killers	Lionsgate	45	1.245333333	11
5	2	animation	4	WALL-E	Disney	89	2.896019067	96
6	2	animation	4	Gnomeo and Juliet	Disney	52	5.387972222	56
7	6	comedy	43	Youth in Revolt	The Weinstein Company	52	1.09	68
8	6	comedy	43	The Heartbreak Kid	Paramount	41	2.129444167	30
9	6	comedy	43	Valentine's Day	Warner Bros.	54	4.184038462	17
10	6	comedy	43	The Ugly Truth	Independent	68	5.402631579	14