# Project 2
# Domain-Specific Data Engineering
### Retail Inventory Management and Sales Forecasting

## Step 1: Define the Scope

1. **Business Objective:** Optimize inventory management and sales forecasting.
2. **Data Sources:**
    - Sales transactions (customer purchases, product details, quantities, timestamps)
    - Inventory logs (stock levels, product movements, dates)
    - Customer data (demographics, purchase history, preferences)
3. **Success Metrics/Goal:** Optimize inventory levels to reduce stockouts by 20% within the next 12 months, while improving sales forecast accuracy by 15%.

## Step 2: Design the Data Pipeline:

1. **Data Ingestion:**

- **Sales transactions: Through file transfer via Vnet to onprem network (VPN) for batch processing (on-premises to Azure SQL data base)**
- **Inventory logs: Copy activity, source: Synapse Warehouse, sink: Azure SQL Database**
- **Customer data: Through Database connectors, Notebook Activity, Source: Azure database for MySQL (Azure data studio), sink: Azure SQL database**

2. **Data Storage: Azure SQL Database**
3. **Data Transformation: Pyspark (notebook)**
4. **Data Warehousing and Analytics: Snowflake, Power BI**

**Step 3. Implementation and Monitoring: Adf monitor**

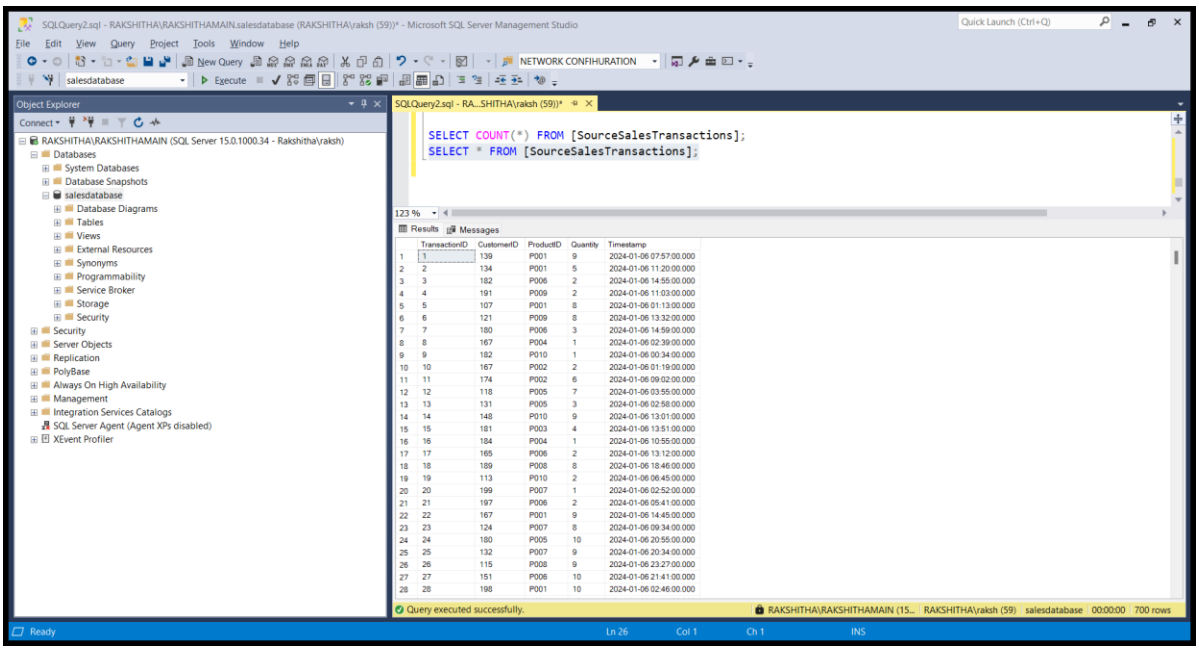
**Step 4: Advanced Considerations:**

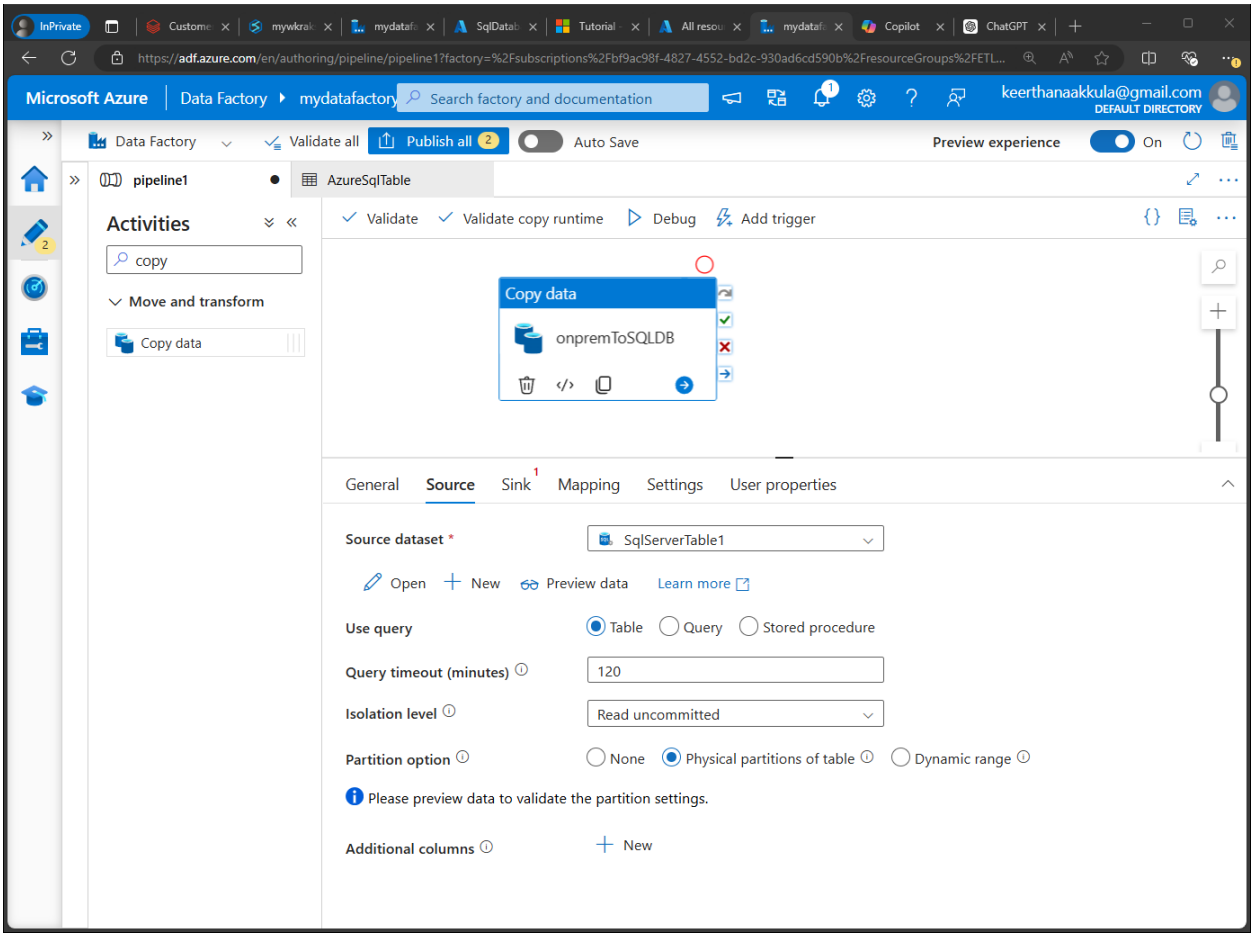
1. **Data Format Exploration:** (CSV vs. Parquet) comparisons in data bricks
2. **Data Pipeline Automation:** No real time data implemented in this project
3. **Medallion Architecture:** Implemented in data bricks
4. **Distribution Indexing:** Used indexing and table distribution on the source tables

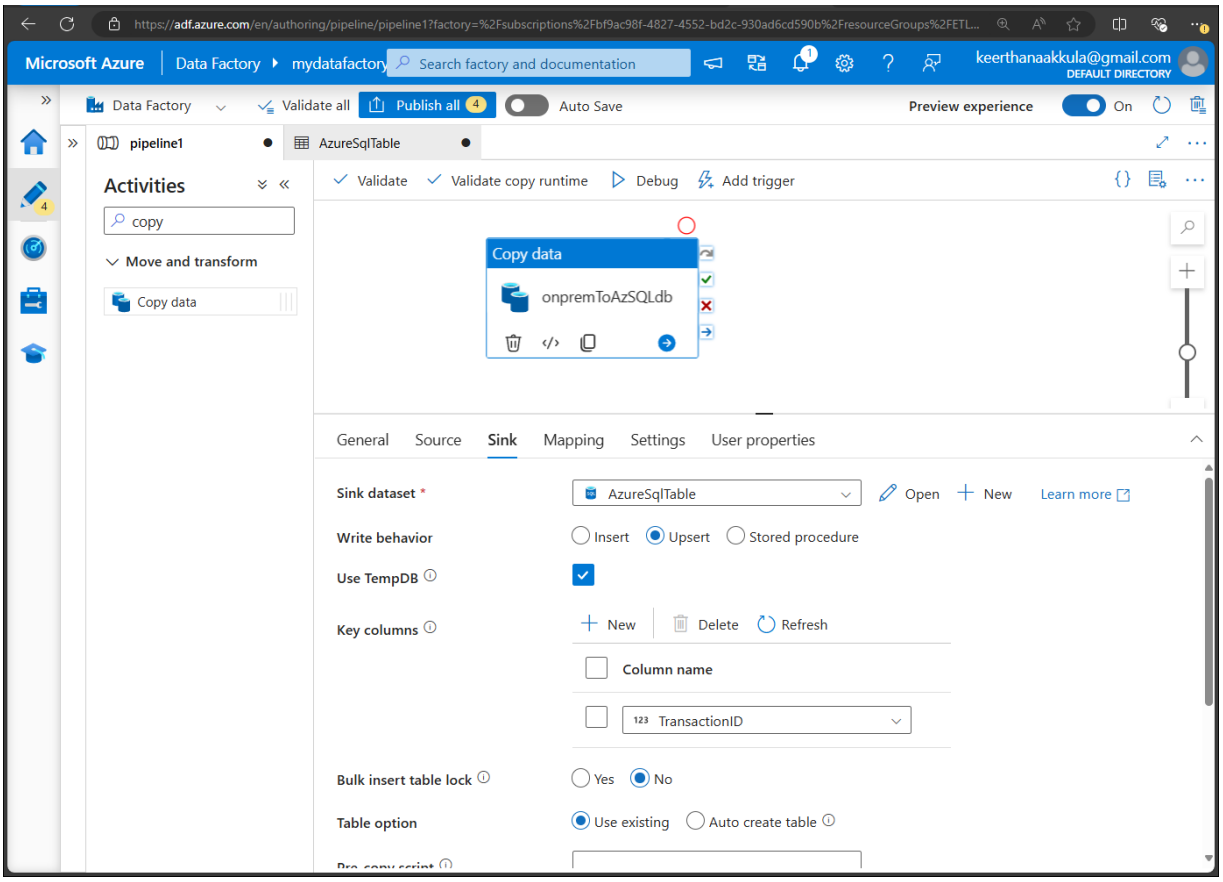
**Implementation of project:**

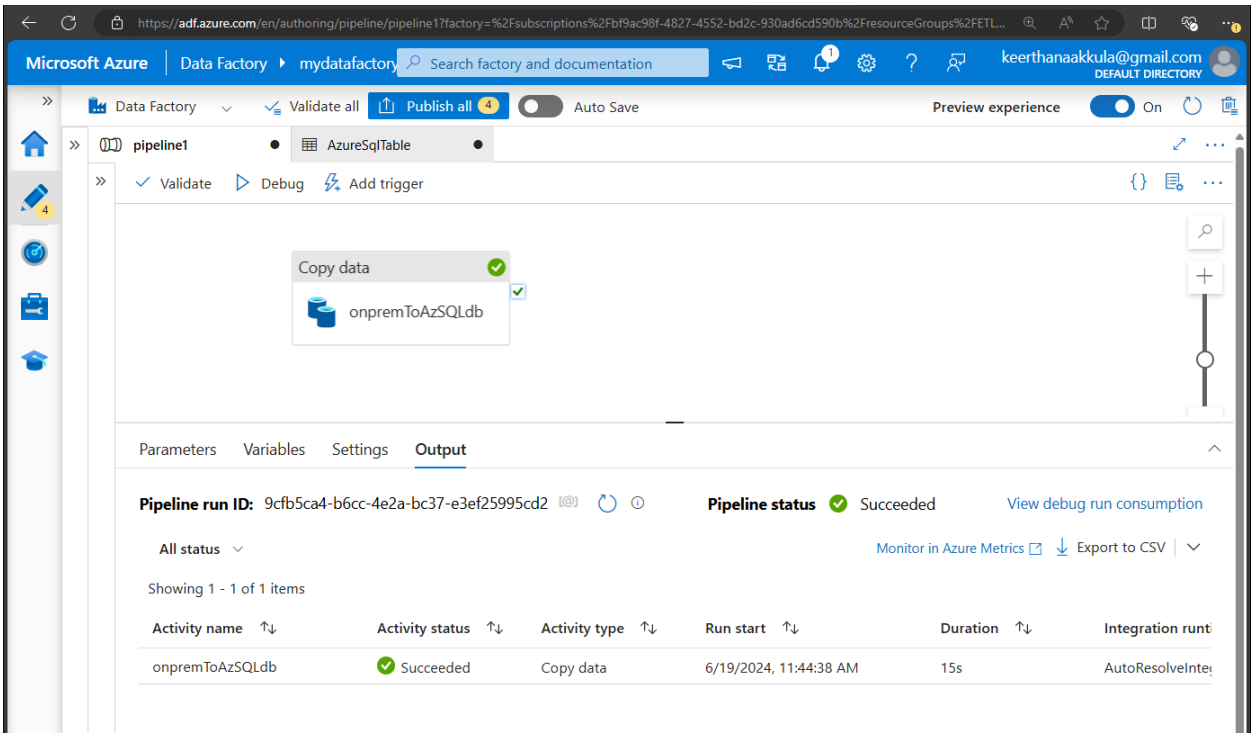
1. **Data Ingestion:**

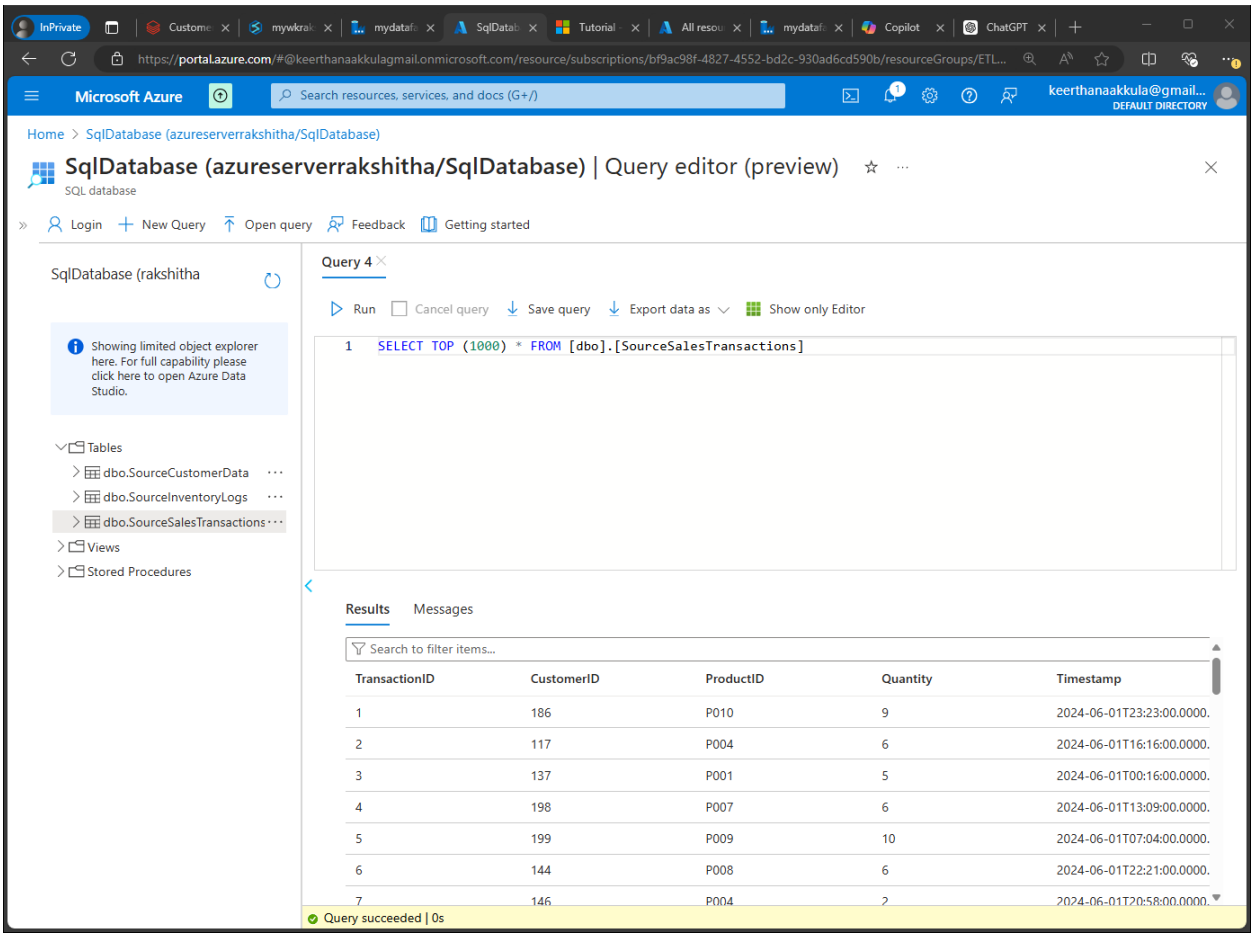
- **Sales transactions: Through file transfer via Vnet to onprem network (VPN) for batch processing (on-premises to Azure SQL data base)**

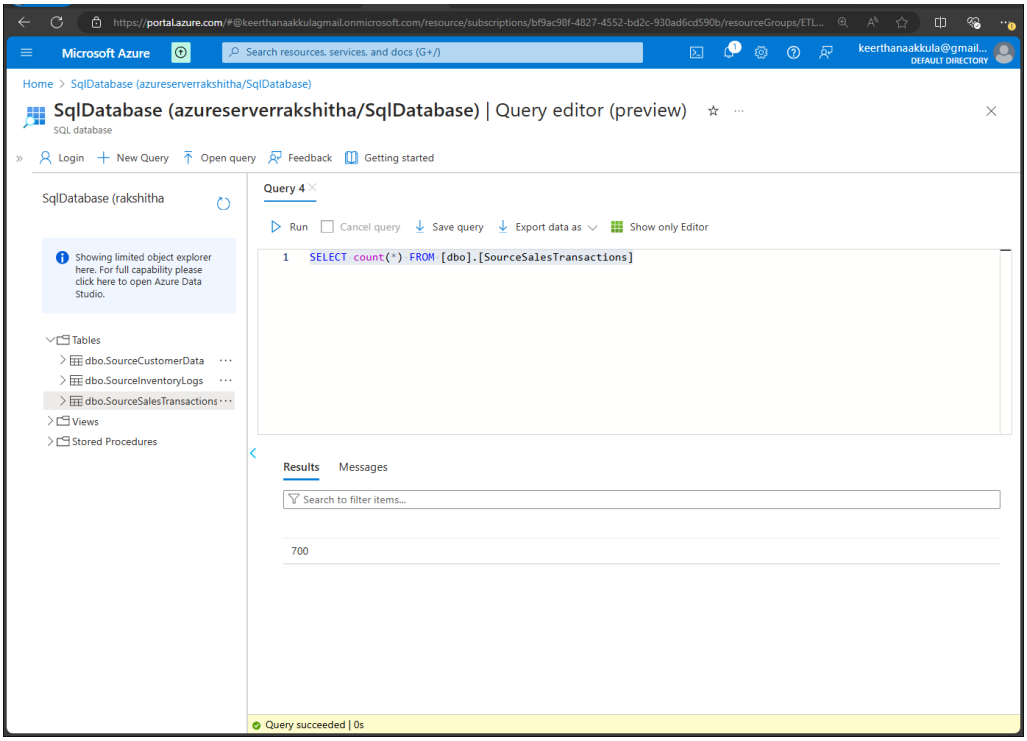

SELECT COUNT(*) FROM [SourceSalesTransactions];
SELECT * FROM [SourceSalesTransactions];

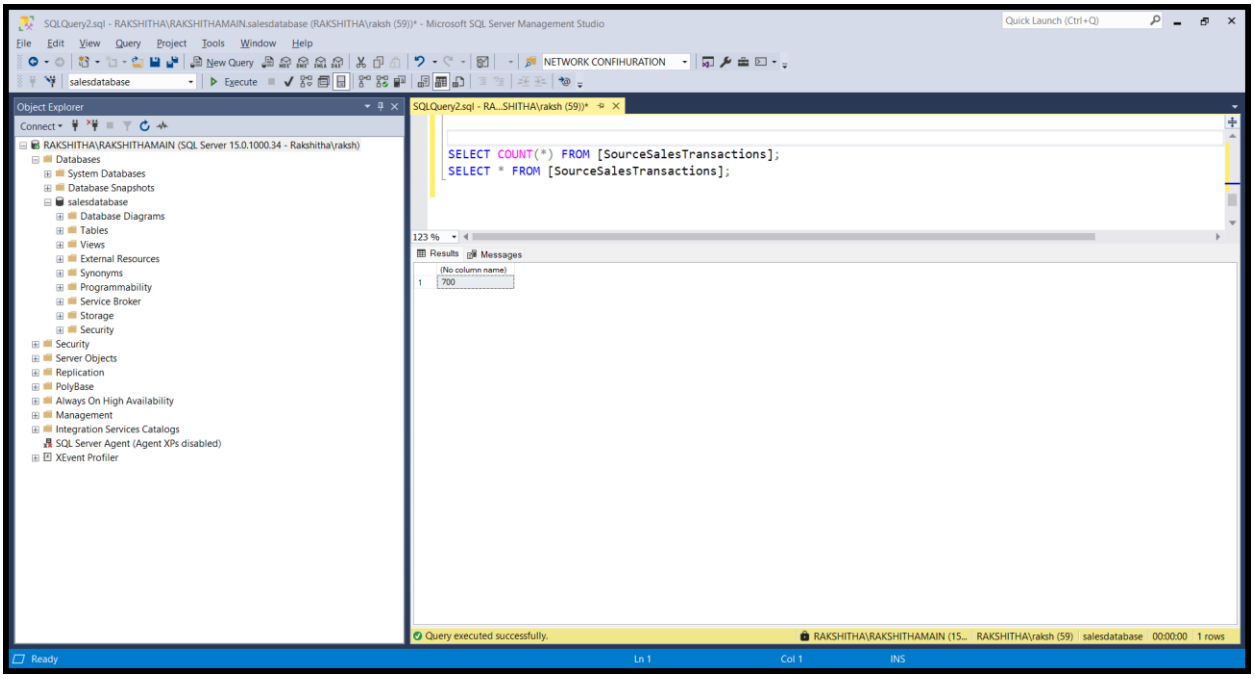

Microsoft Azure
Data Factory
mydatafactory
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all 2
Auto Save
Preview experience
On
pipeline1
AzureSqlTable
Activities
copy
Move and transform
Copy data
Validate
Validate copy runtime
Debug
Add trigger
Copy data
onpremToSQLDB
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
SqlServerTable1
Open
New
Preview data
Learn more
Use query
Table
Query
Stored procedure
Query timeout (minutes)
120
Isolation level
Read uncommitted
Partition option
None
Physical partitions of table
Dynamic range
Please preview data to validate the partition settings.
Additional columns
New


Microsoft Azure
Data Factory
mydatafactory
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 4
Auto Save
Preview experience
On
pipeline1
AzureSqlTable
Activities
copy
Move and transform
Copy data
Validate
Validate copy runtime
Debug
Add trigger
Copy data
onpremToAzSQLdb
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
AzureSqlTable
Open
New
Learn more
Write behavior
Insert
Upsert
Stored procedure
Use TempDB
Key columns
New
Delete
Refresh
Column name
TransactionID
Bulk insert table lock
Yes
No
Table option
Use existing
Auto create table

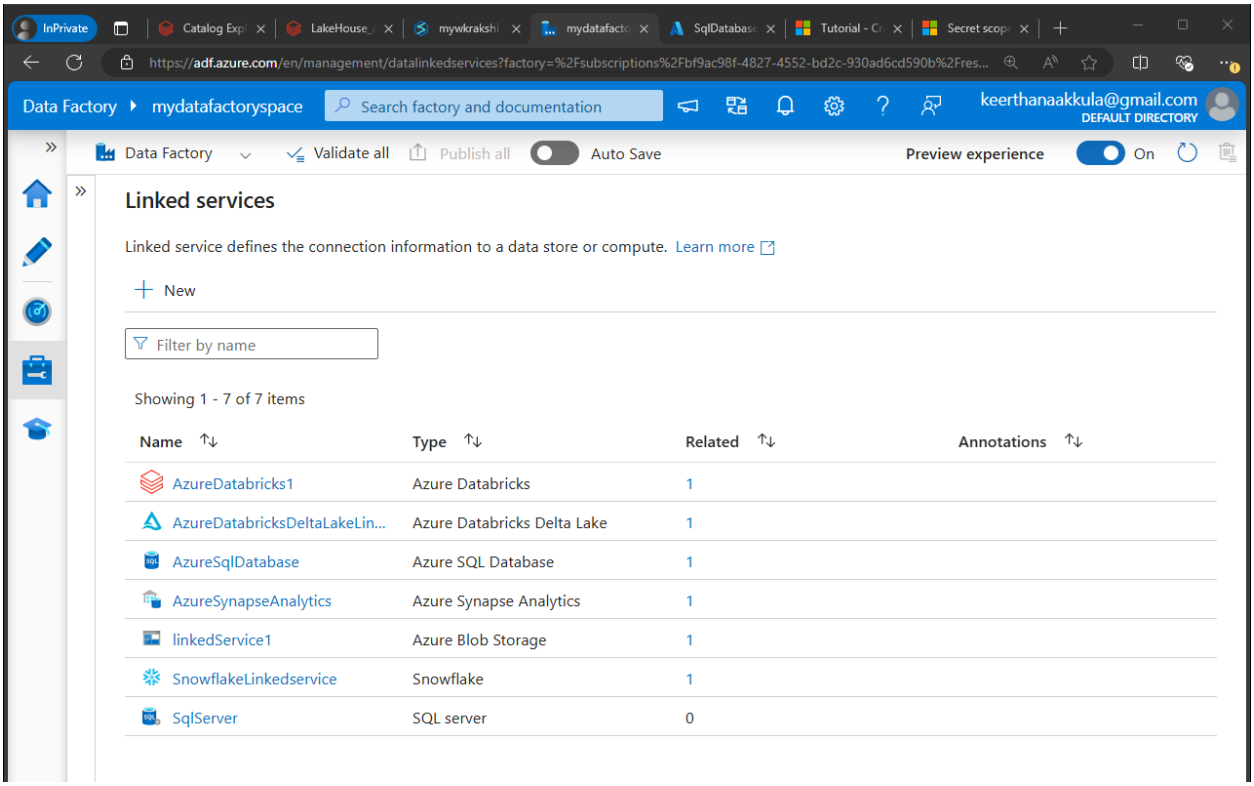

Microsoft Azure
Data Factory
mydatafactory
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 4
Auto Save
Preview experience
On
pipeline1
AzureSqlTable
Validate
Debug
Add trigger
Copy data
onpremToAzSQLdb
Parameters
Variables
Settings
Output
Pipeline run ID: 9cfb5ca4-b6cc-4e2a-bc37-e3ef25995cd2
Pipeline status
Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runti
onpremToAzSQLdb
Succeeded
Copy data
6/19/2024, 11:44:38 AM
15s
AutoResolveInte

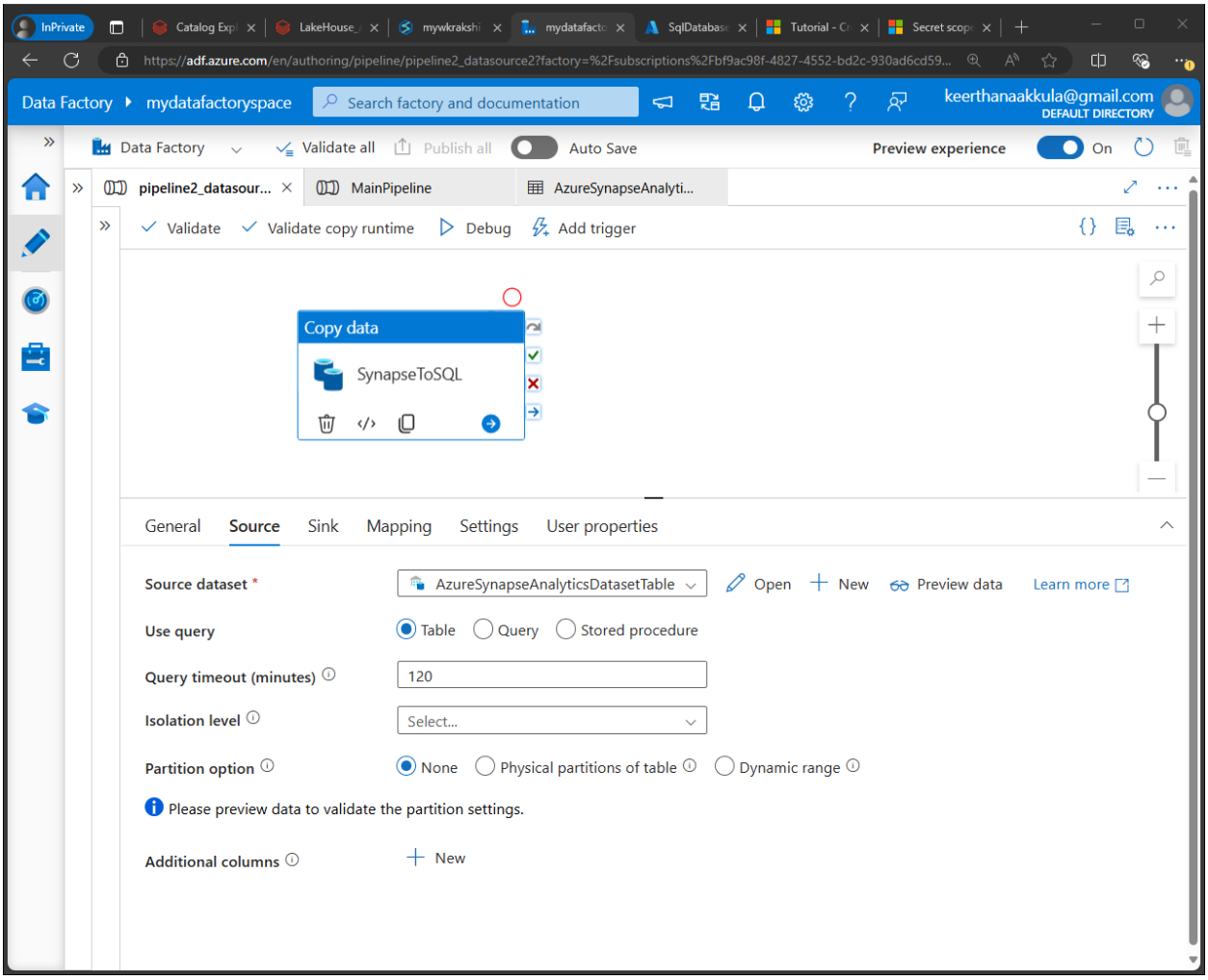

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > SqlDatabase (azureserverrakshitha/SqlDatabase)
SqlDatabase (azureserverrakshitha/SqlDatabase) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Getting started
SqlDatabase (rakshitha
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.SourceCustomerData
dbo.SourceInventoryLogs
dbo.SourceSalesTransactions
Views
Stored Procedures
Query 4
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT TOP (1000) * FROM [dbo].[SourceSalesTransactions]
Results
Messages
Search to filter items...
TransactionID
CustomerID
ProductID
Quantity
Timestamp
1 186 P010 9 2024-06-01T23:23:00.0000.
2 117 P004 6 2024-06-01T16:16:00.0000.
3 137 P001 5 2024-06-01T00:16:00.0000.
4 198 P007 6 2024-06-01T13:09:00.0000.
5 199 P009 10 2024-06-01T07:04:00.0000.
6 144 P008 6 2024-06-01T22:21:00.0000.
7 146 P004 2 2024-06-01T20:58:00.0000.
Query succeeded | 0s

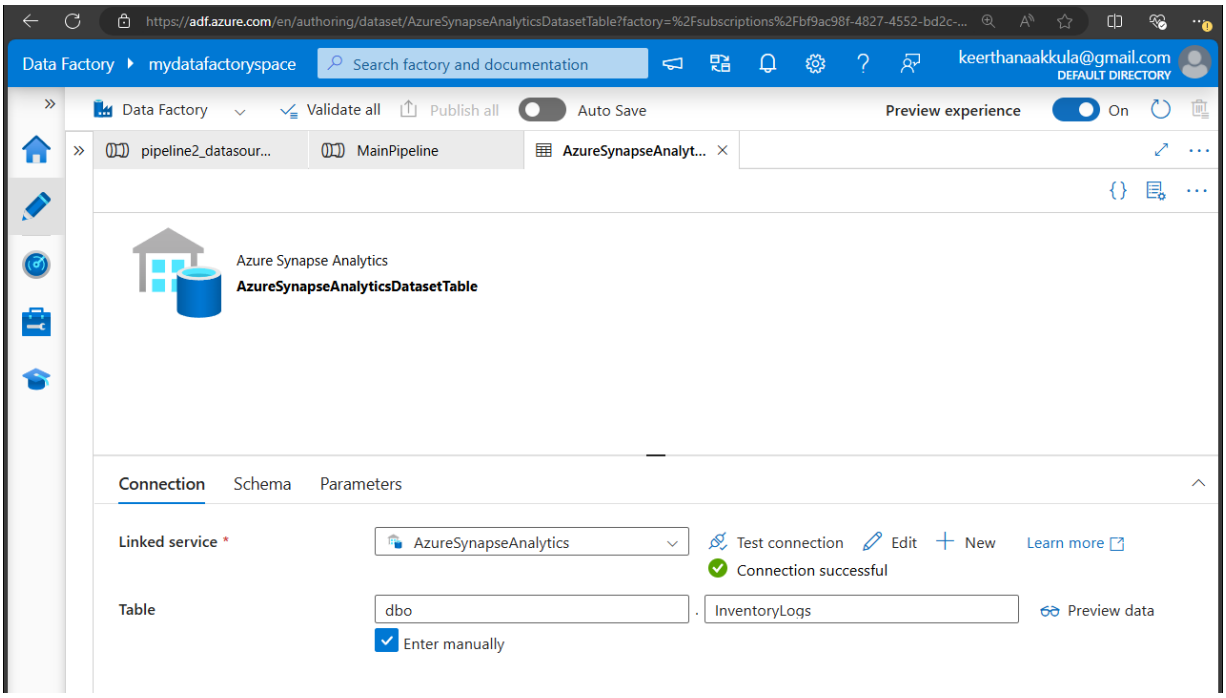

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY
Home > SqlDatabase (azureserverrakshitha/SqlDatabase)
SqlDatabase (azureserverrakshitha/SqlDatabase) | Query editor (preview)
SQL database
Login
New Query
Open query
Feedback
Getting started
SqlDatabase (rakshitha
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.SourceCustomerData
dbo.SourceInventoryLogs
dbo.SourceSalesTransactions
Views
Stored Procedures
Query 4
Run
Cancel query
Save query
Export data as
Show only Editor
1 SELECT count(*) FROM [dbo].[SourceSalesTransactions]
Results
Messages
Search to filter items...
700
Query succeeded | 0s

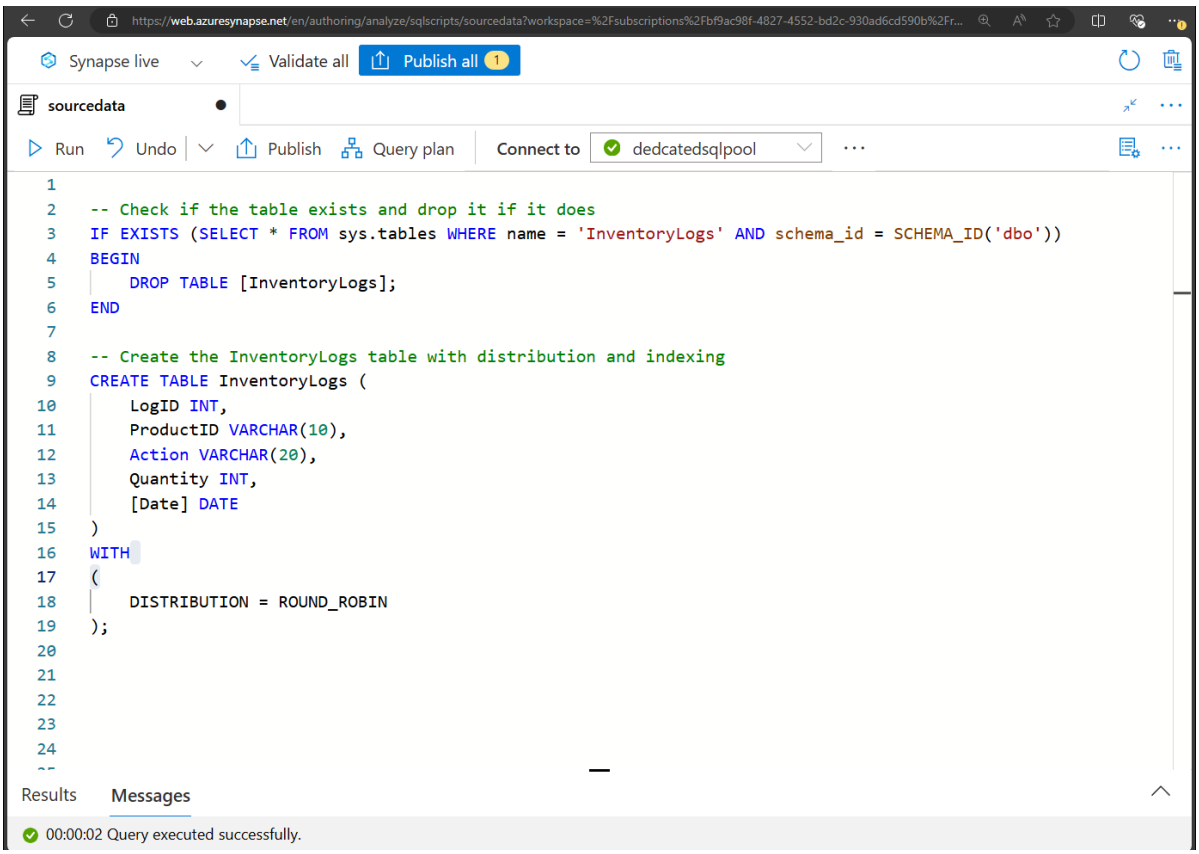


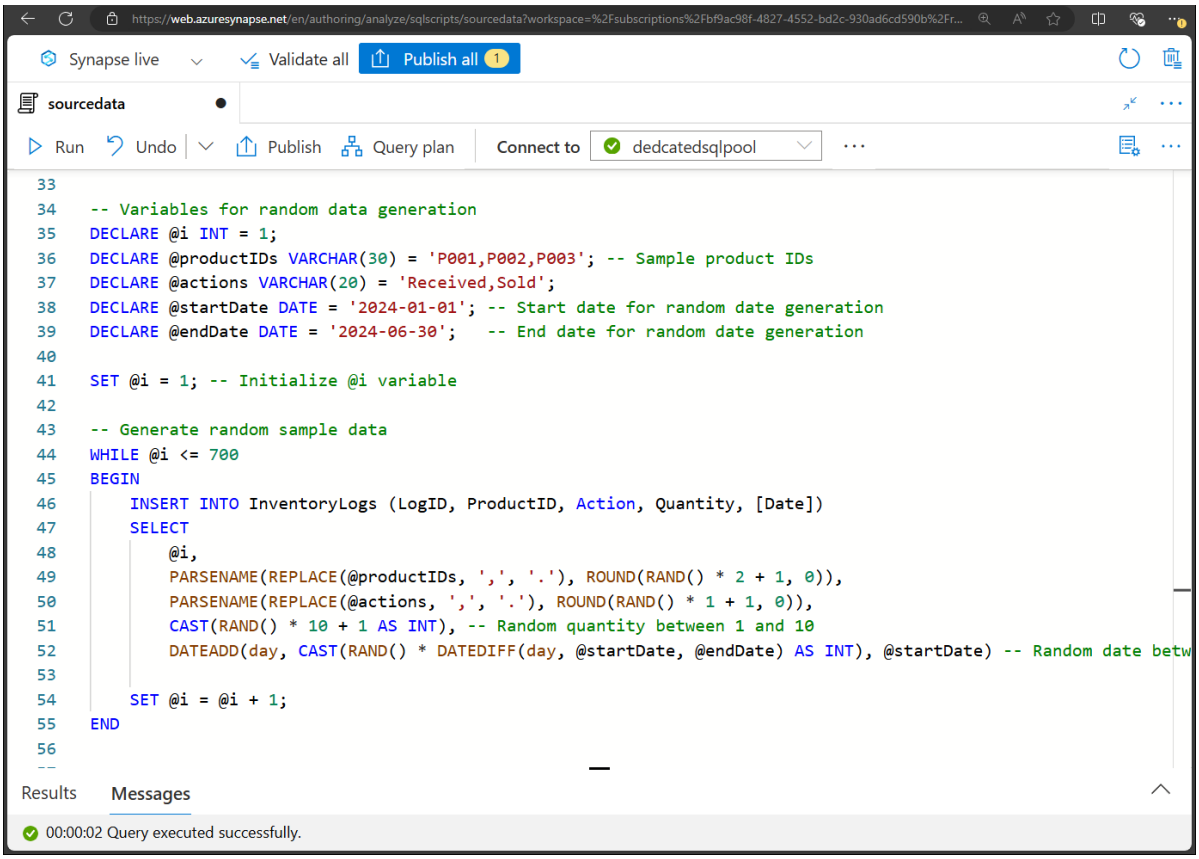

- **Inventory logs: Copy activity, source: Synapse Warehouse, sink: Azure SQL Database**

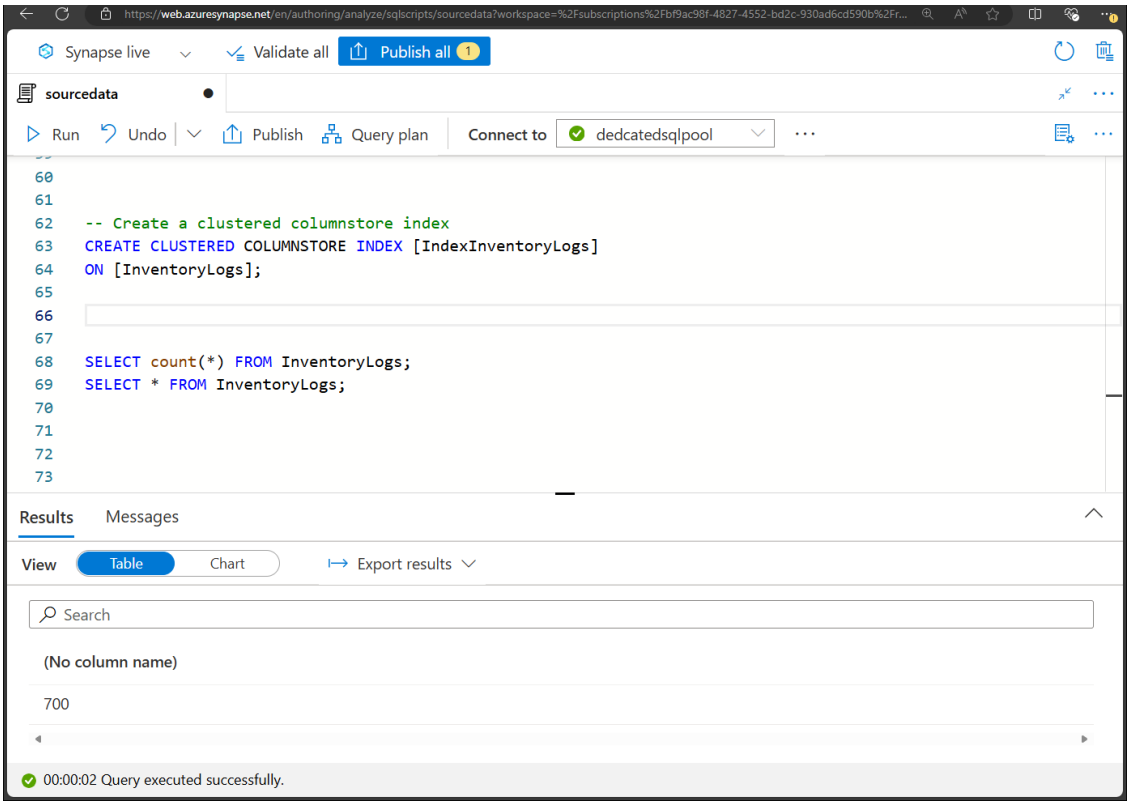


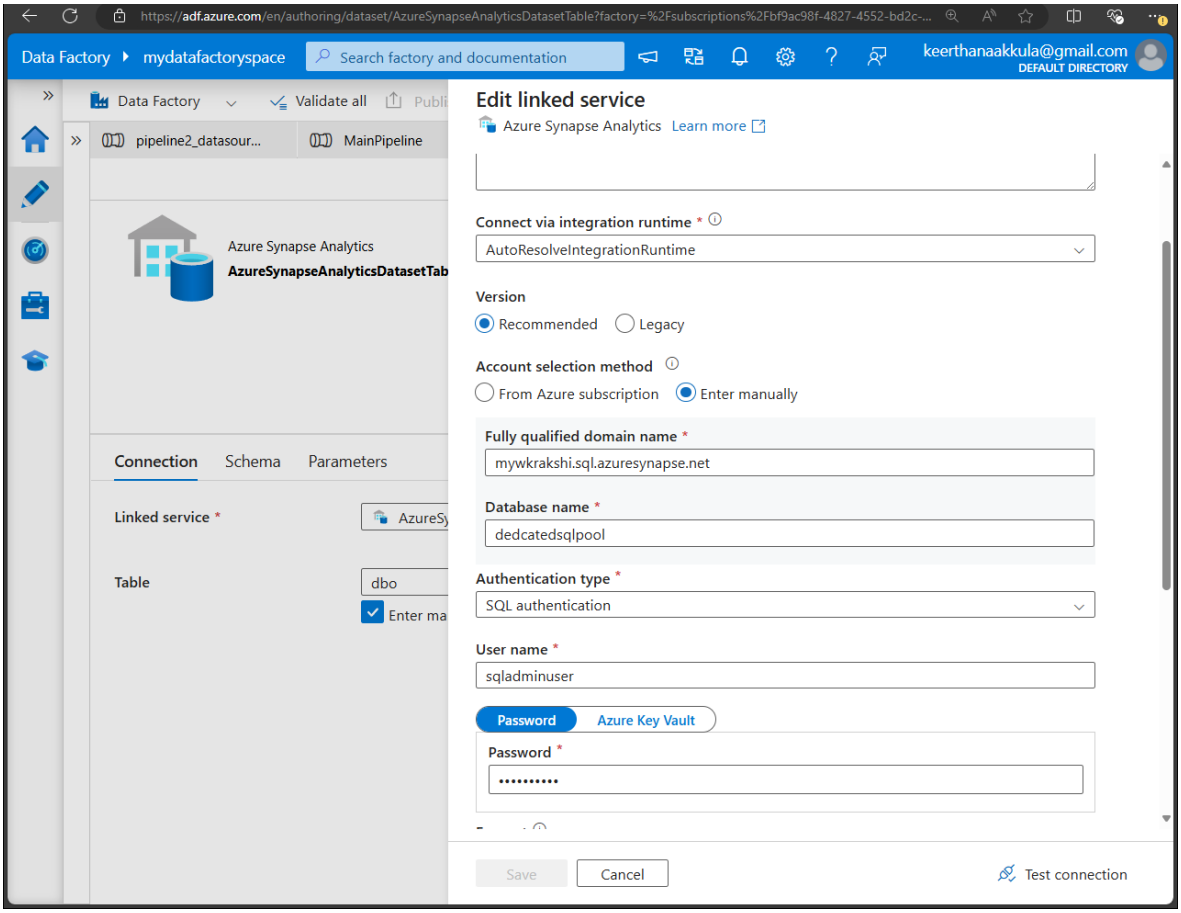

Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Auto Save
Preview experience
On
pipeline2_datasour...
MainPipeline
AzureSynapseAnalyti...
Validate
Validate copy runtime
Debug
Add trigger
Copy data
SynapseToSQL
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
AzureSynapseAnalyticsDatasetTable
Open
New
Preview data
Learn more
Use query
Table
Query
Stored procedure
Query timeout (minutes)
120
Isolation level
Select...
Partition option
None
Physical partitions of table
Dynamic range
Please preview data to validate the partition settings.
Additional columns
New

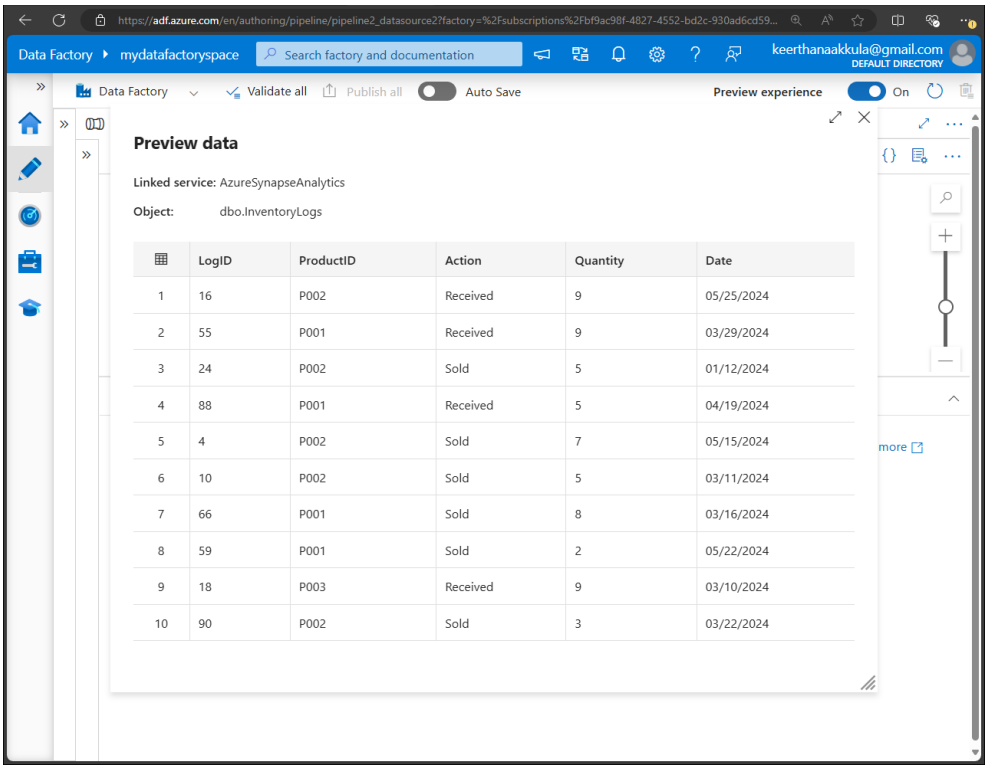

Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Auto Save
Preview experience
On
pipeline2_datasour...
MainPipeline
AzureSynapseAnalyt...
Azure Synapse Analytics
AzureSynapseAnalyticsDatasetTable
Connection
Schema
Parameters
Linked service *
AzureSynapseAnalytics
Test connection
Edit
New
Learn more
Connection successful
Table
dbo
InventoryLogs
Preview data
Enter manually


Synapse live
Validate all
Publish all 1
sourcedata
Run
Undo
Publish
Query plan
Connect to
dedcatedsqlpool
1 -- Check if the table exists and drop it if it does
2 IF EXISTS (SELECT * FROM sys.tables WHERE name = 'InventoryLogs' AND schema_id = SCHEMA_ID('dbo'))
3 BEGIN
4 DROP TABLE [InventoryLogs];
5 END
6
7 -- Create the InventoryLogs table with distribution and indexing
8 CREATE TABLE InventoryLogs (
9 LogID INT,
10 ProductID VARCHAR(10),
11 Action VARCHAR(20),
12 Quantity INT,
13 [Date] DATE
14 )
15 WITH
16 (
17 DISTRIBUTION = ROUND_ROBIN
18 );
19
20
21
22
23
24
Results
Messages
00:00:02 Query executed successfully.

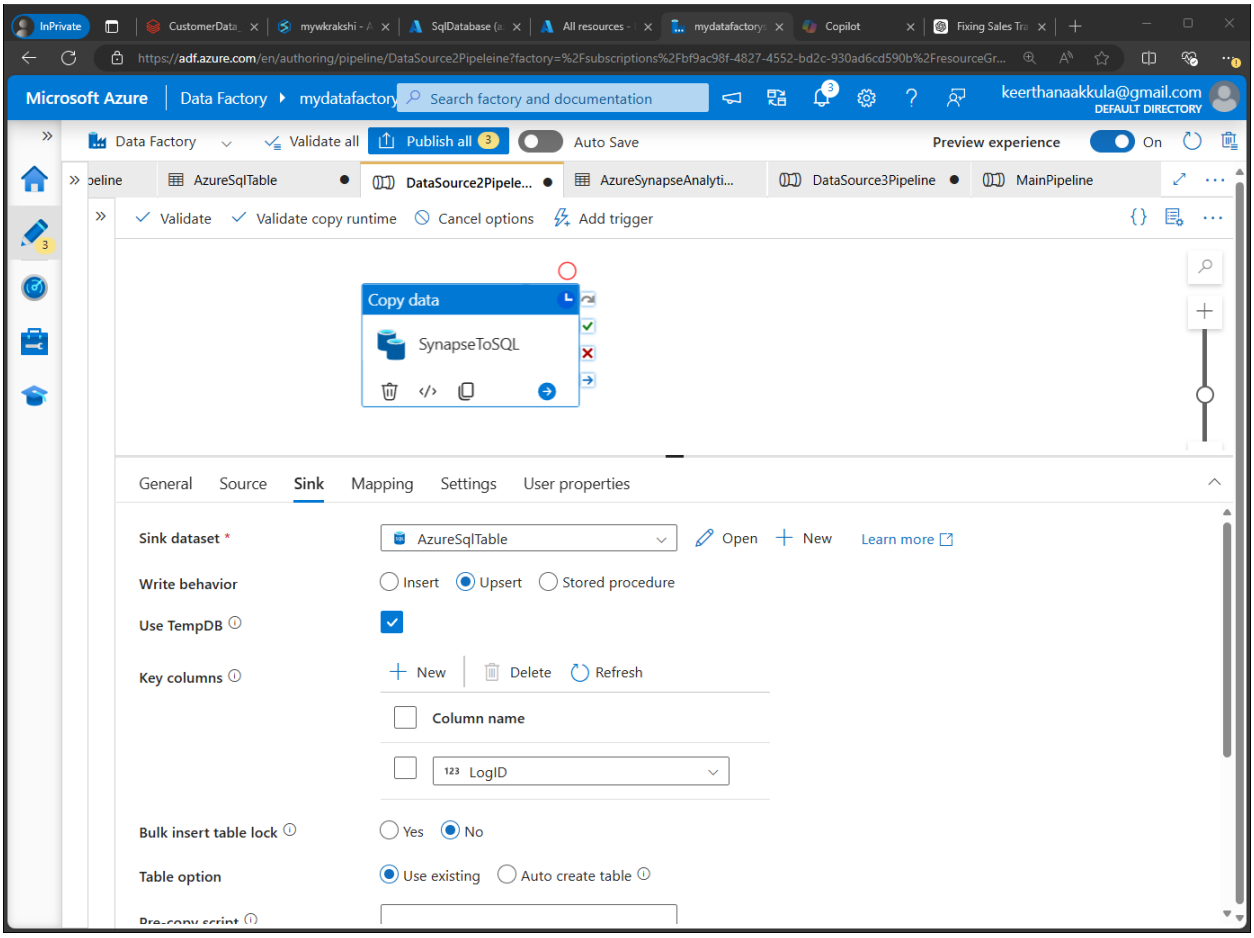

Synapse live
Validate all
Publish all 1
sourcedata
Run
Undo
Publish
Query plan
Connect to
dedcatedsqlpool
33
34 -- Variables for random data generation
35 DECLARE @i INT = 1;
36 DECLARE @productIDs VARCHAR(30) = 'P001,P002,P003'; -- Sample product IDs
37 DECLARE @actions VARCHAR(20) = 'Received,Sold';
38 DECLARE @startDate DATE = '2024-01-01'; -- Start date for random date generation
39 DECLARE @endDate DATE = '2024-06-30'; -- End date for random date generation
40
41 SET @i = 1; -- Initialize @i variable
42
43 -- Generate random sample data
44 WHILE @i <= 700
45 BEGIN
46 INSERT INTO InventoryLogs (LogID, ProductID, Action, Quantity, [Date])
47 SELECT
48 @i,
49 PARSENAME(REPLACE(@productIDs, ',', '.'), ROUND(RAND() * 2 + 1, 0)),
50 PARSENAME(REPLACE(@actions, ',', '.'), ROUND(RAND() * 1 + 1, 0)),
51 CAST(RAND() * 10 + 1 AS INT), -- Random quantity between 1 and 10
52 DATEADD(day, CAST(RAND() * DATEDIFF(day, @startDate, @endDate) AS INT), @startDate) -- Random date betw
53
54 SET @i = @i + 1;
55 END
56
Results
Messages
00:00:02 Query executed successfully.

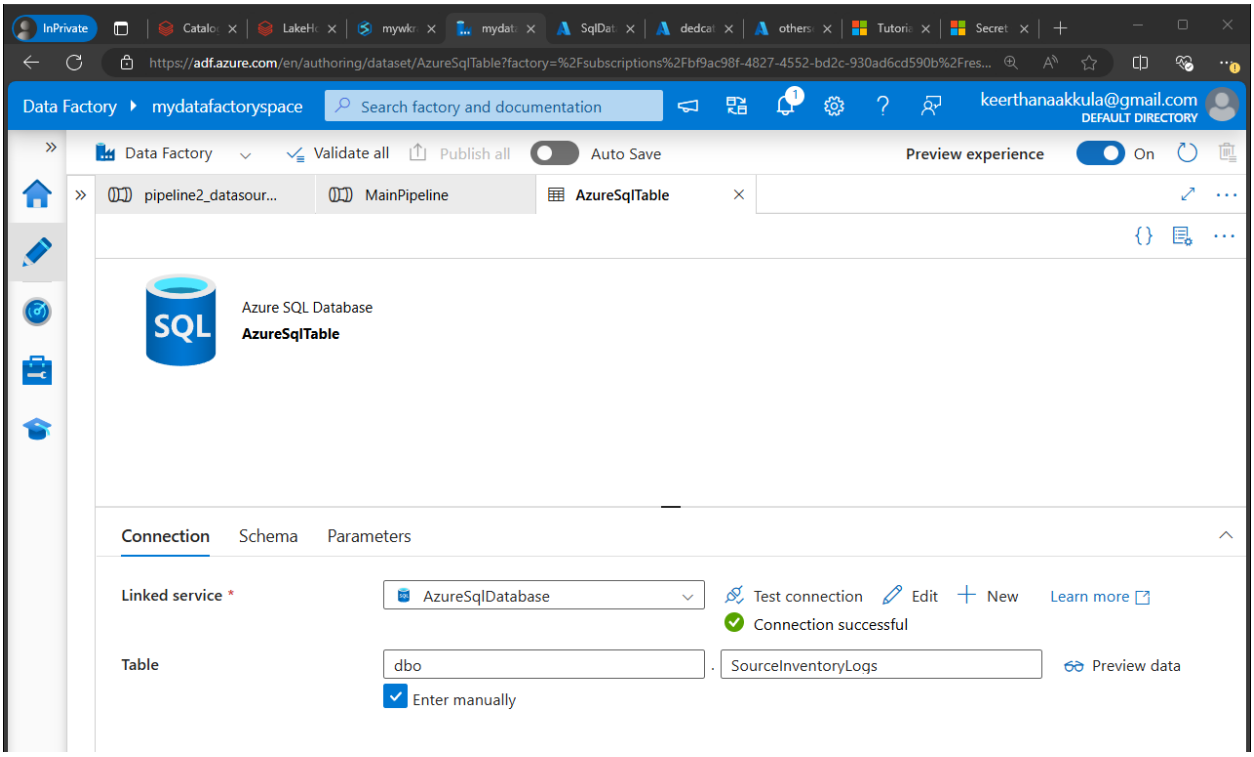

Synapse live
Validate all
Publish all 1
sourcedata
Run
Undo
Publish
Query plan
Connect to
dedcatedsqlpool
60
61
62 -- Create a clustered columnstore index
63 CREATE CLUSTERED COLUMNSTORE INDEX [IndexInventoryLogs]
64 ON [InventoryLogs];
65
66
67
68 SELECT count(*) FROM InventoryLogs;
69 SELECT * FROM InventoryLogs;
70
71
72
73
Results
Messages
View
Table
Chart
Export results
Search
(No column name)
700
00:00:02 Query executed successfully.

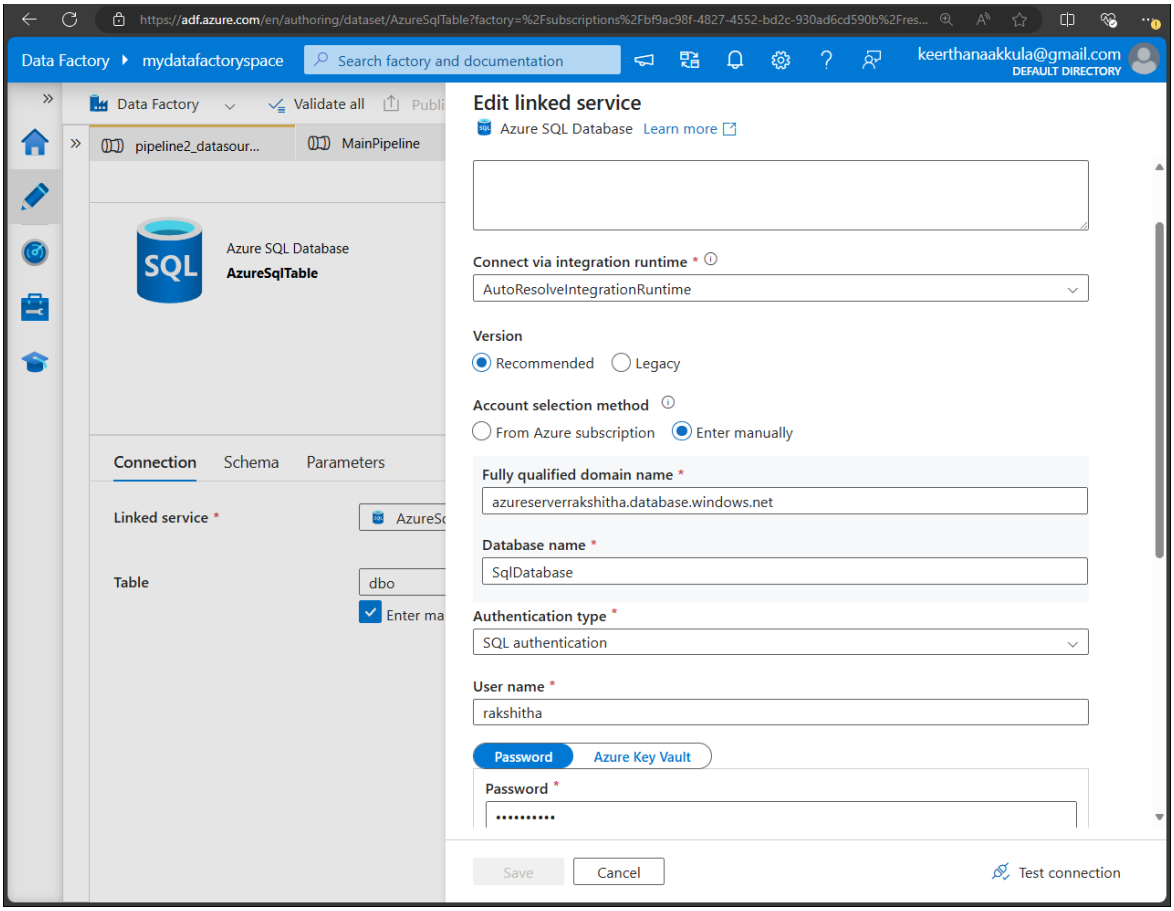

Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Edit linked service
Azure Synapse Analytics
Learn more
Connect via integration runtime *
AutoResolveIntegrationRuntime
Version
Recommended
Legacy
Account selection method
From Azure subscription
Enter manually
Fully qualified domain name *
mywkrakshi.sql.azuresynapse.net
Database name *
dedcatedsqlpool
Authentication type *
SQL authentication
User name *
sqladminuser
Password
Azure Key Vault
Password *
Connection
Schema
Parameters
Linked service *
Table
dbo
Save
Cancel
Test connection

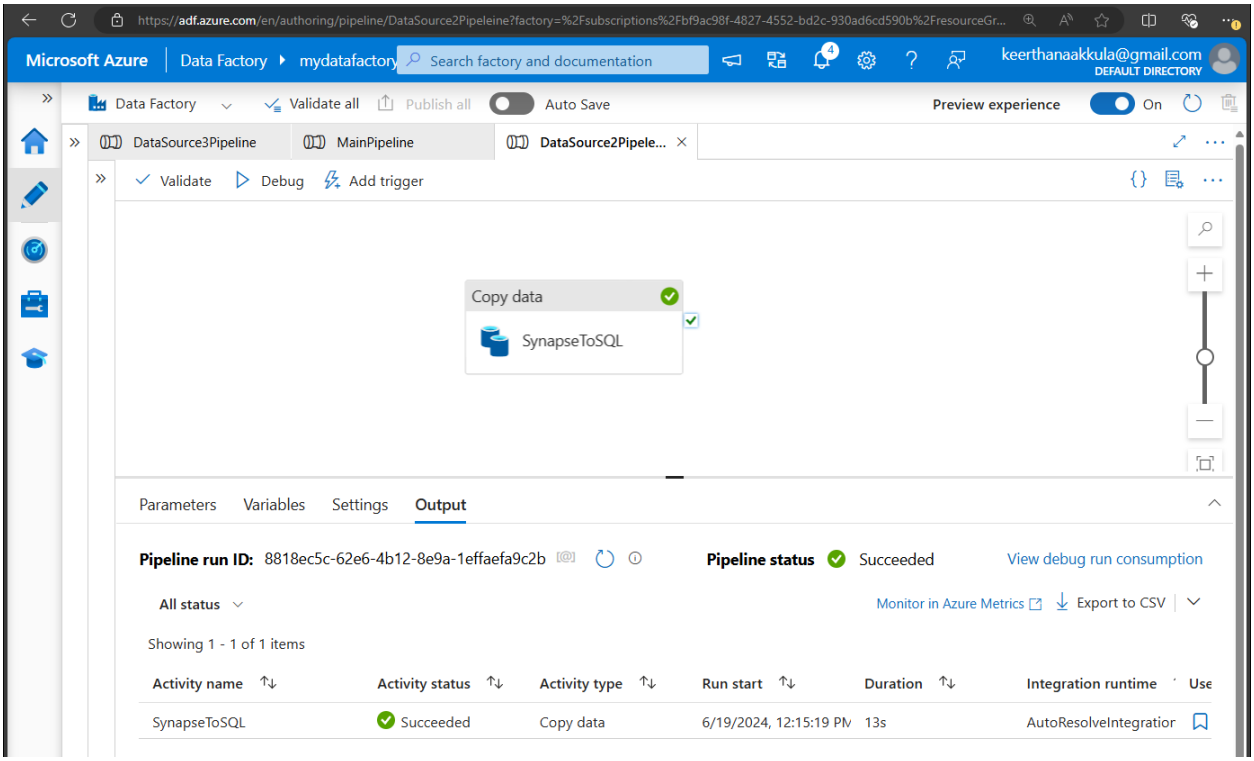


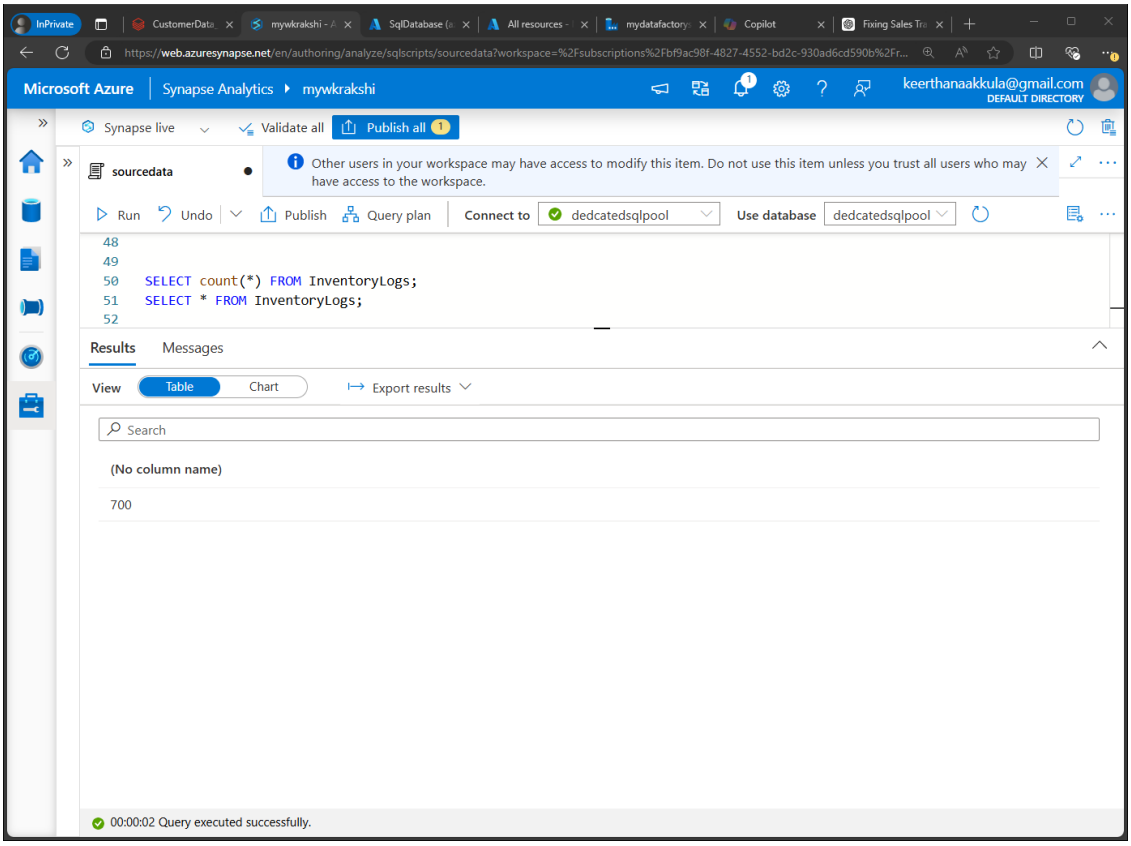

| | LogID | ProductID | Action | Quantity | Date |
|---|---|---|---|---|---|
| 1 | 16 | P002 | Received | 9 | 05/25/2024 |
| 2 | 55 | P001 | Received | 9 | 03/29/2024 |
| 3 | 24 | P002 | Sold | 5 | 01/12/2024 |
| 4 | 88 | P001 | Received | 5 | 04/19/2024 |
| 5 | 4 | P002 | Sold | 7 | 05/15/2024 |
| 6 | 10 | P002 | Sold | 5 | 03/11/2024 |
| 7 | 66 | P001 | Sold | 8 | 03/16/2024 |
| 8 | 59 | P001 | Sold | 2 | 05/22/2024 |
| 9 | 18 | P003 | Received | 9 | 03/10/2024 |
| 10 | 90 | P002 | Sold | 3 | 03/22/2024 |

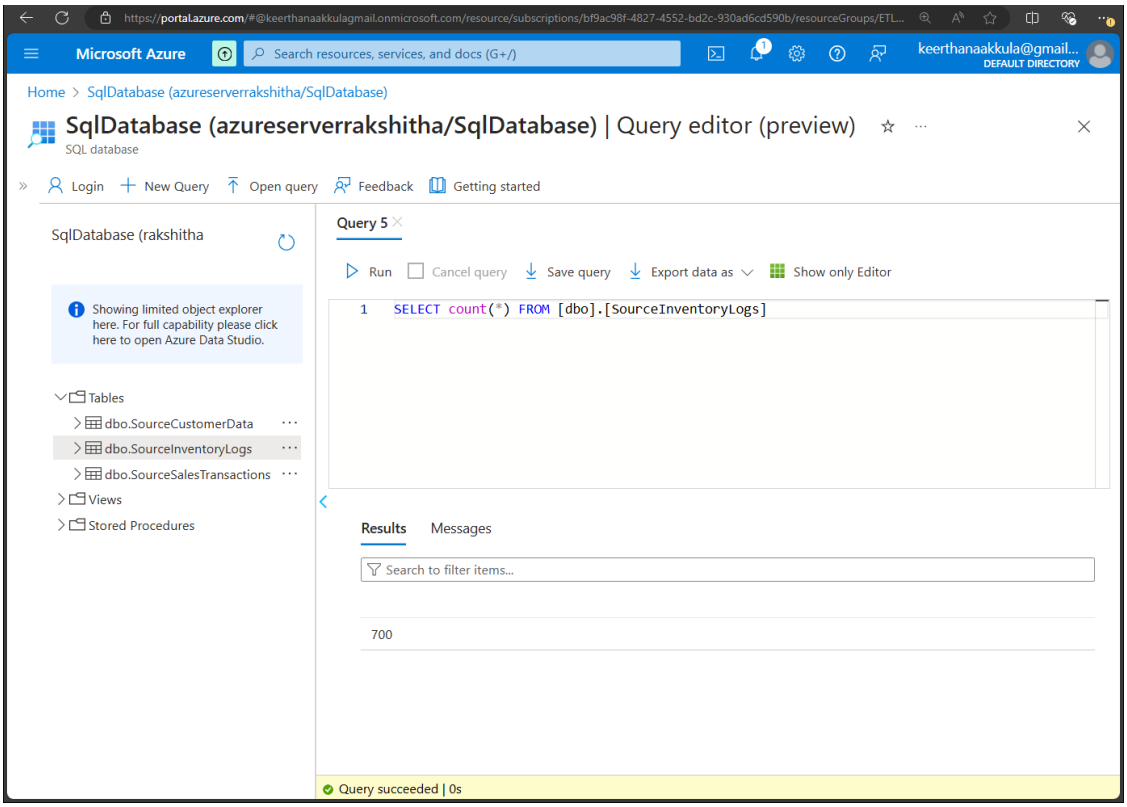



Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Auto Save
Preview experience
On
pipeline2_datasour...
MainPipeline
AzureSqlTable
Azure SQL Database
AzureSqlTable
Connection
Schema
Parameters
Linked service *
AzureSqlDatabase
Test connection
Edit
New
Learn more
Connection successful
Table
dbo
SourceInventoryLogs
Preview data
Enter manually


Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Edit linked service
Azure SQL Database
Learn more
Connect via integration runtime *
AutoResolveIntegrationRuntime
Version
Recommended
Legacy
Account selection method
From Azure subscription
Enter manually
Fully qualified domain name *
azureserverrakshitha.database.windows.net
Database name *
SqlDatabase
Authentication type *
SQL authentication
User name *
rakshitha
Password
Azure Key Vault
Password *
Save
Cancel
Test connection


Microsoft Azure
Data Factory
mydatafactory
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
MainPipeline
DataSource2Pipele...
Validate
Debug
Add trigger
Copy data
SynapseToSQL
Parameters
Variables
Settings
Output
Pipeline run ID: 8818ec5c-62e6-4b12-8e9a-1effaefa9c2b
Pipeline status Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
SynapseToSQL
Succeeded
Copy data
6/19/2024, 12:15:19 PM
13s
AutoResolveIntegration


Microsoft Azure
Synapse Analytics
mywkrakshi
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Synapse live
Validate all
Publish all
sourcedata
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.
Run
Undo
Publish
Query plan
Connect to
dedcatedsqlpool
Use database
dedcatedsqlpool
48
49
50 SELECT count(*) FROM InventoryLogs;
51 SELECT * FROM InventoryLogs;
52
Results
Messages
View
Table
Chart
Export results
Search
(No column name)
700
00:00:02 Query executed successfully.



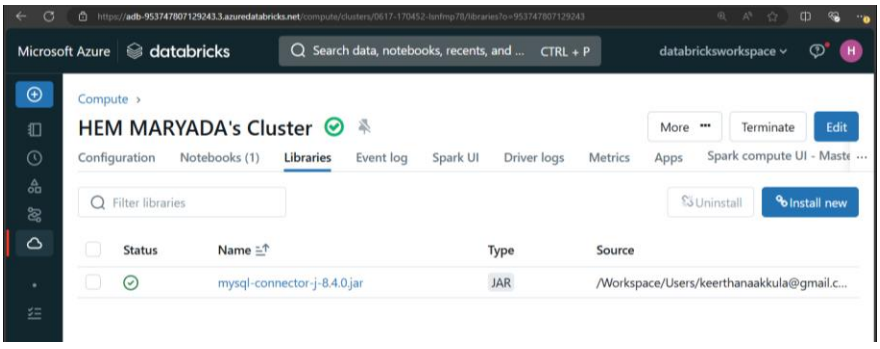

- **Customer data: Through Database connectors, Notebook Activity, Source: Azure database for MySQL (Azure data studio), sink: Azure SQL database**

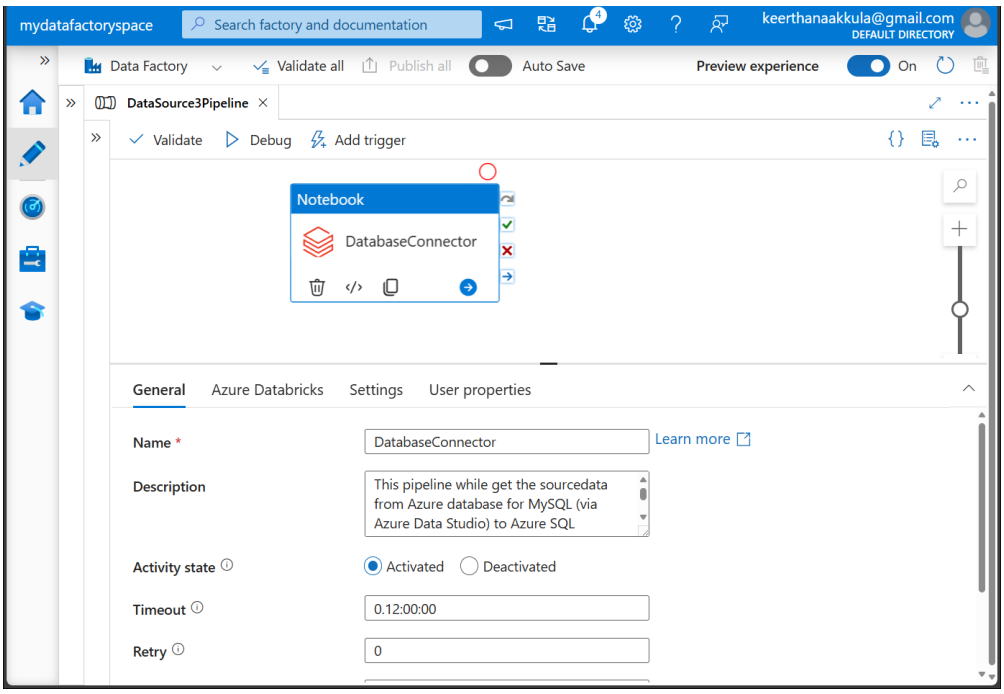
mysql-connector-j-8.4.0.tar.gz

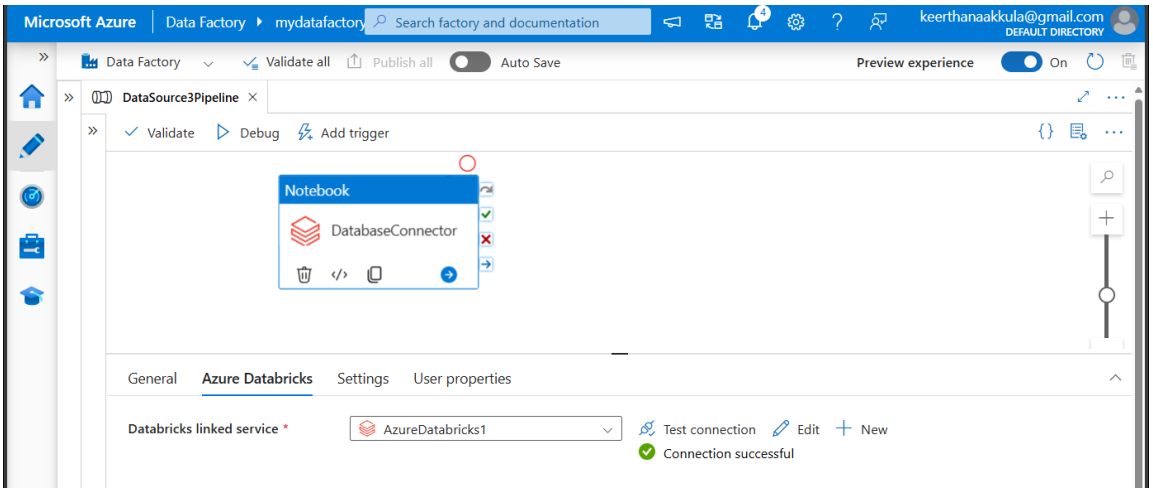
CustomerData_DatabaseConnector.py.html

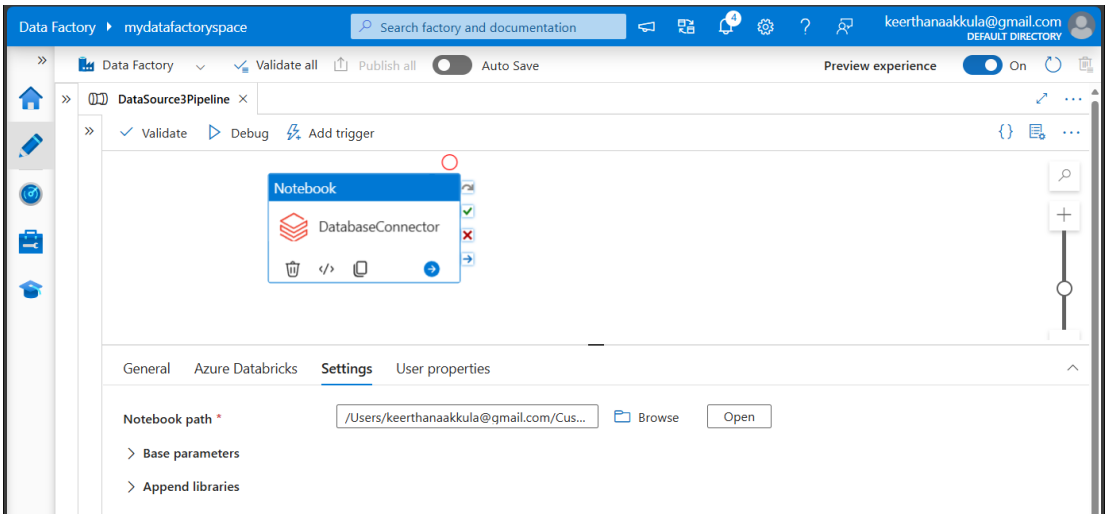
We need to import the mysql connector jar file into databricks and connect to Azure database for MySQL via JDBC connections

To connect to an Azure Database for MySQL from Databricks using JDBC, you need to import the MySQL Connector/J jar file into your Databricks environment. This connector allows Databricks to interact with the MySQL database over JDBC (Java Database Connectivity)




mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
Validate
Debug
Add trigger
Notebook
DatabaseConnector
General
Azure Databricks
Settings
User properties
Name *
DatabaseConnector
Learn more
Description
This pipeline while get the sourcedata from Azure database for MySQL (via Azure Data Studio) to Azure SQL
Activity state
Activated
Deactivated
Timeout
0.12:00:00
Retry
0


Microsoft Azure
Data Factory
mydatafactory
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
Validate
Debug
Add trigger
Notebook
DatabaseConnector
General
Azure Databricks
Settings
User properties
Databricks linked service *
AzureDatabricks1
Test connection
Edit
New
Connection successful


Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
Validate
Debug
Add trigger
Notebook
DatabaseConnector
General
Azure Databricks
Settings
User properties
Notebook path *
/Users/keerthanaakkula@gmail.com/Cus...
Browse
Open
Base parameters
Append libraries

## Edit linked service

Azure Databricks Learn more

AutoResolveIntegrationRuntime

**Account selection method** *

○ From Azure subscription ◉ Enter manually

**Databrick Workspace URL** *

https://adb-953747807129243.3.azuredatabricks.net

**Authentication type** *

Access Token

Access token | Azure Key Vault

**Access token** *

••••••••••

**Select cluster**

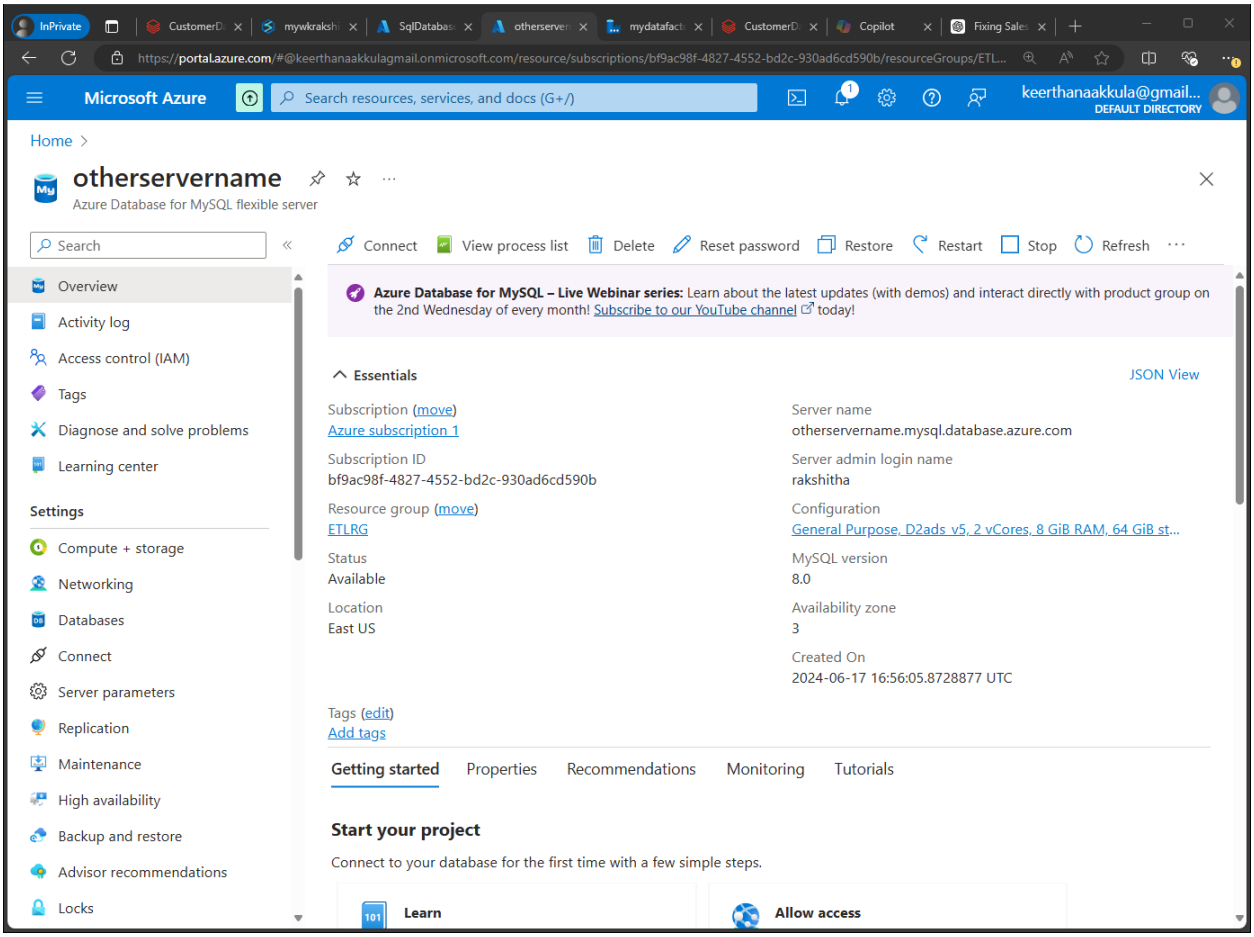
○ New job cluster ◉ Existing interactive cluster ○ Existing instance pool

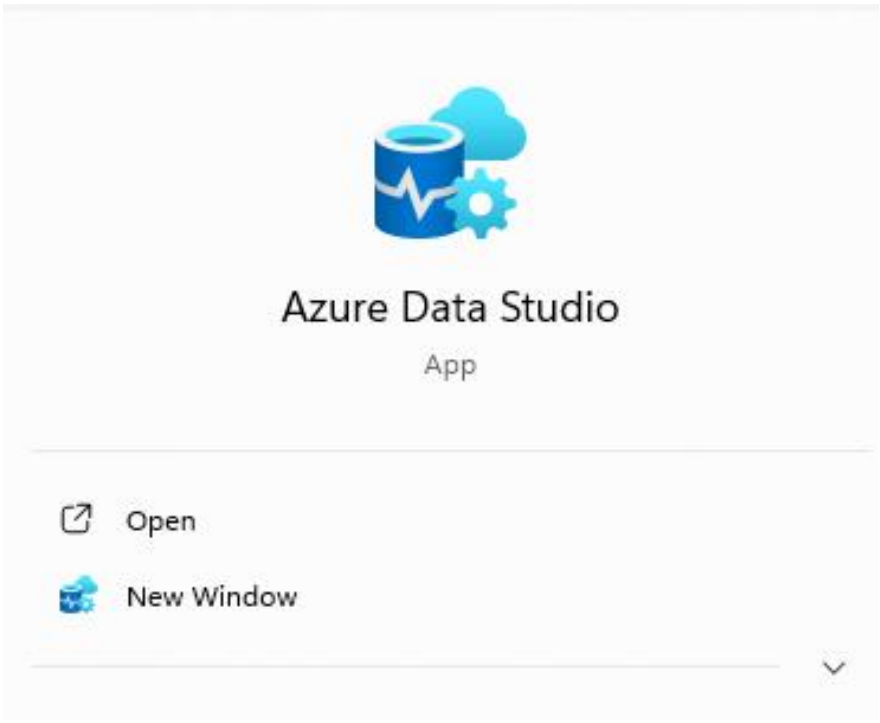
**Existing cluster ID** *

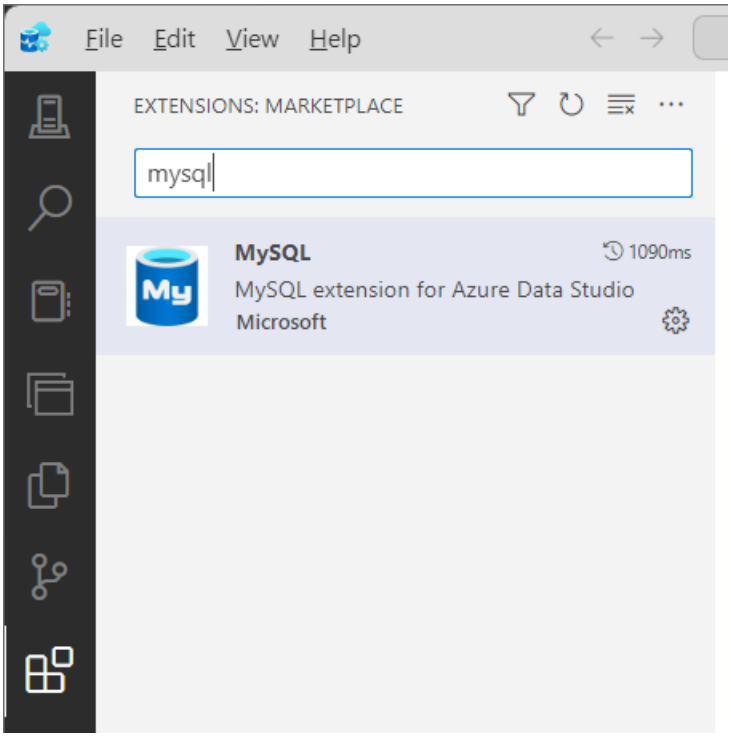
0617-170452-lsnfmp78

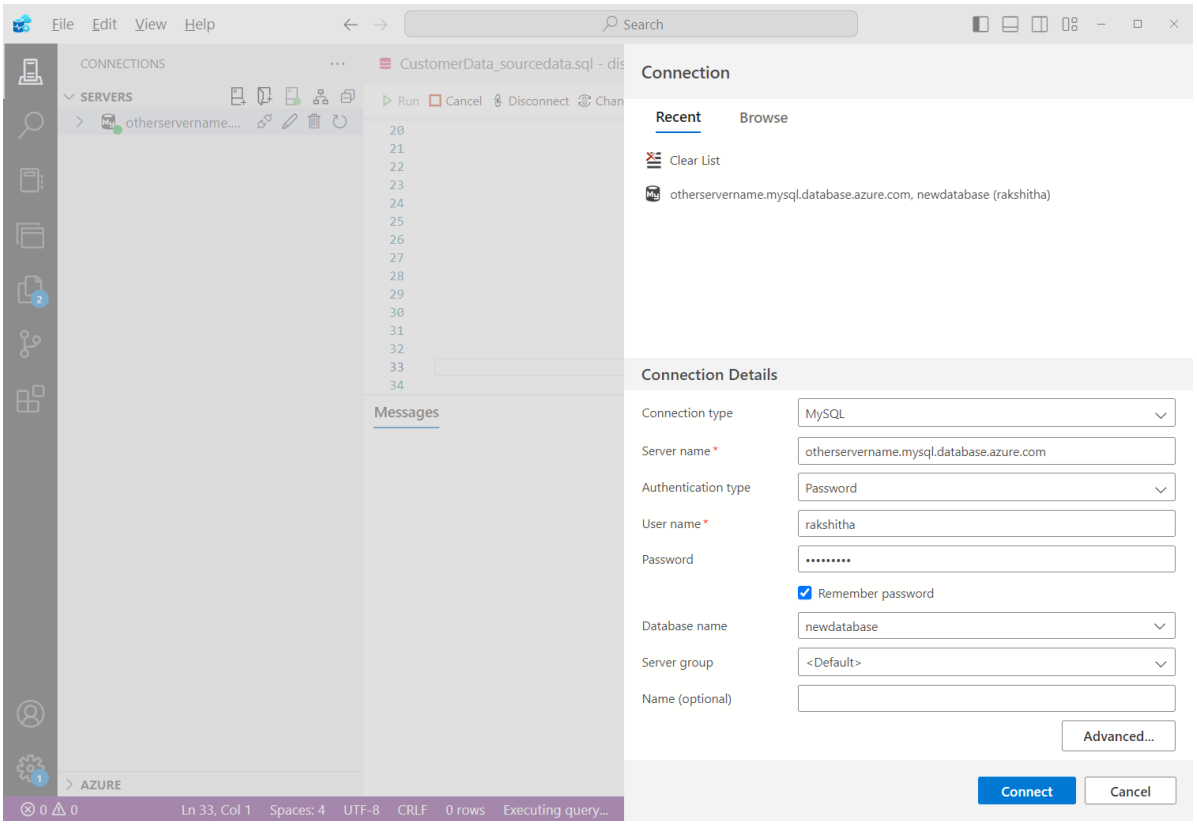
Connection successful

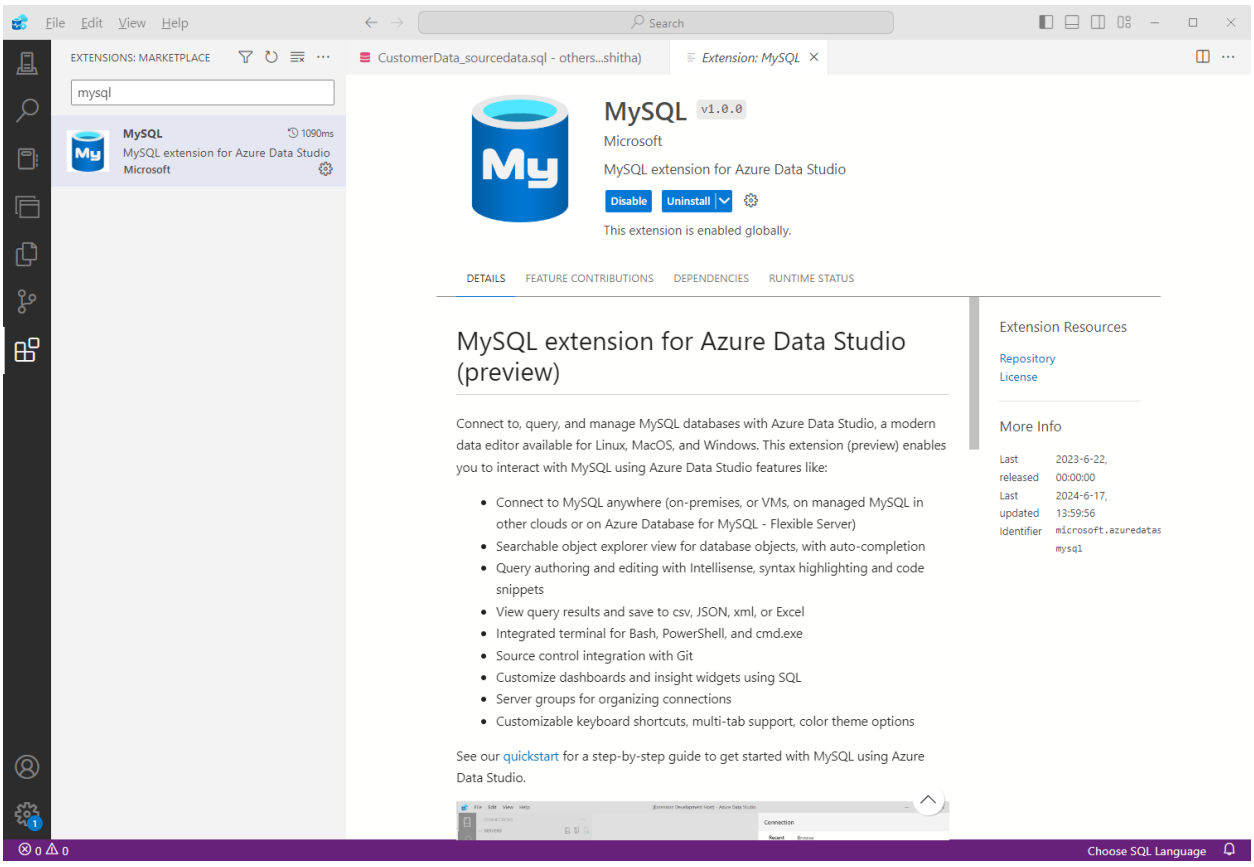
Save | Cancel | Test connection

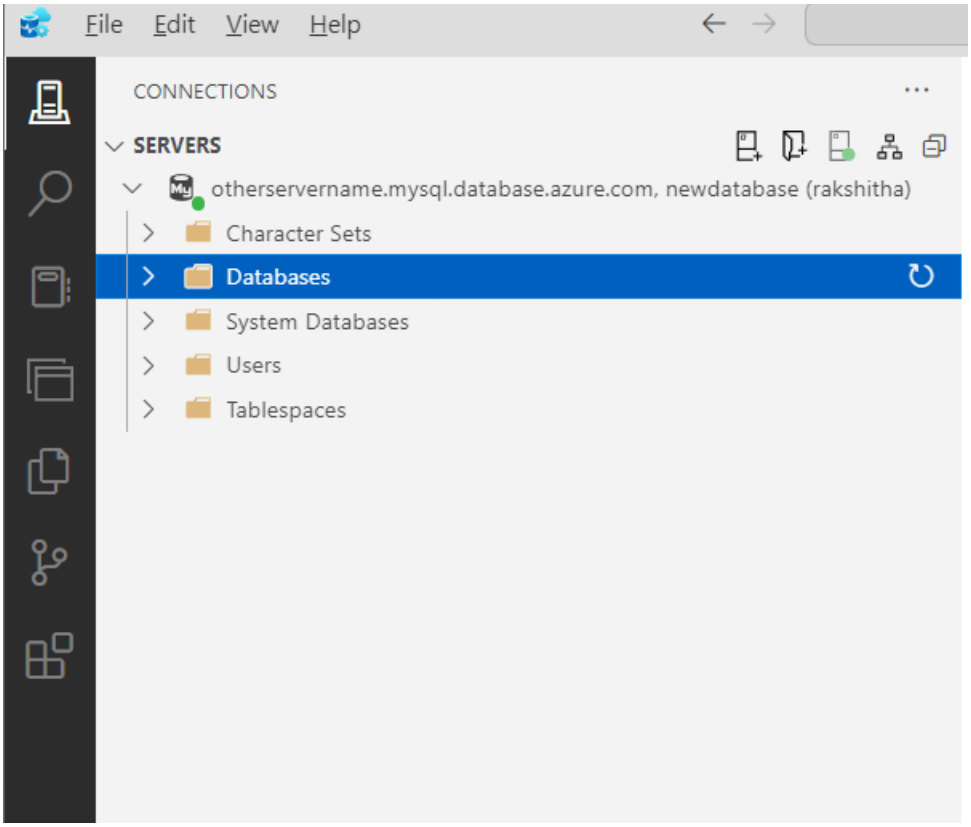


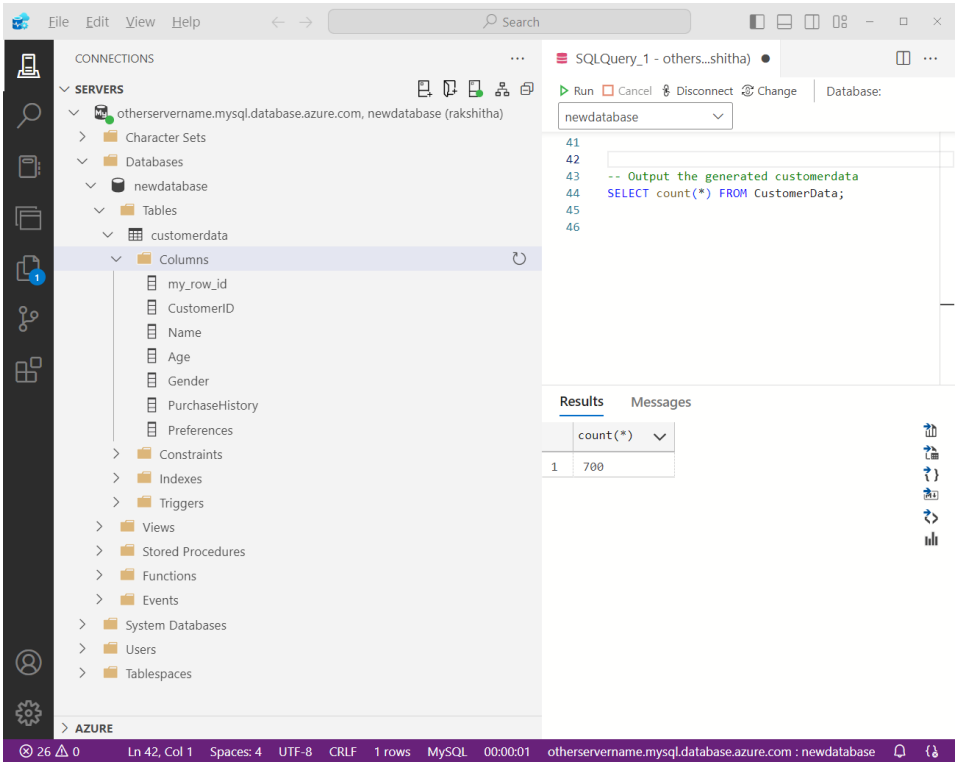

Azure Data Studio
App
Open
New Window


File
Edit
View
Help
EXTENSIONS: MARKETPLACE
mysql
MySQL
1090ms
MySQL extension for Azure Data Studio
Microsoft

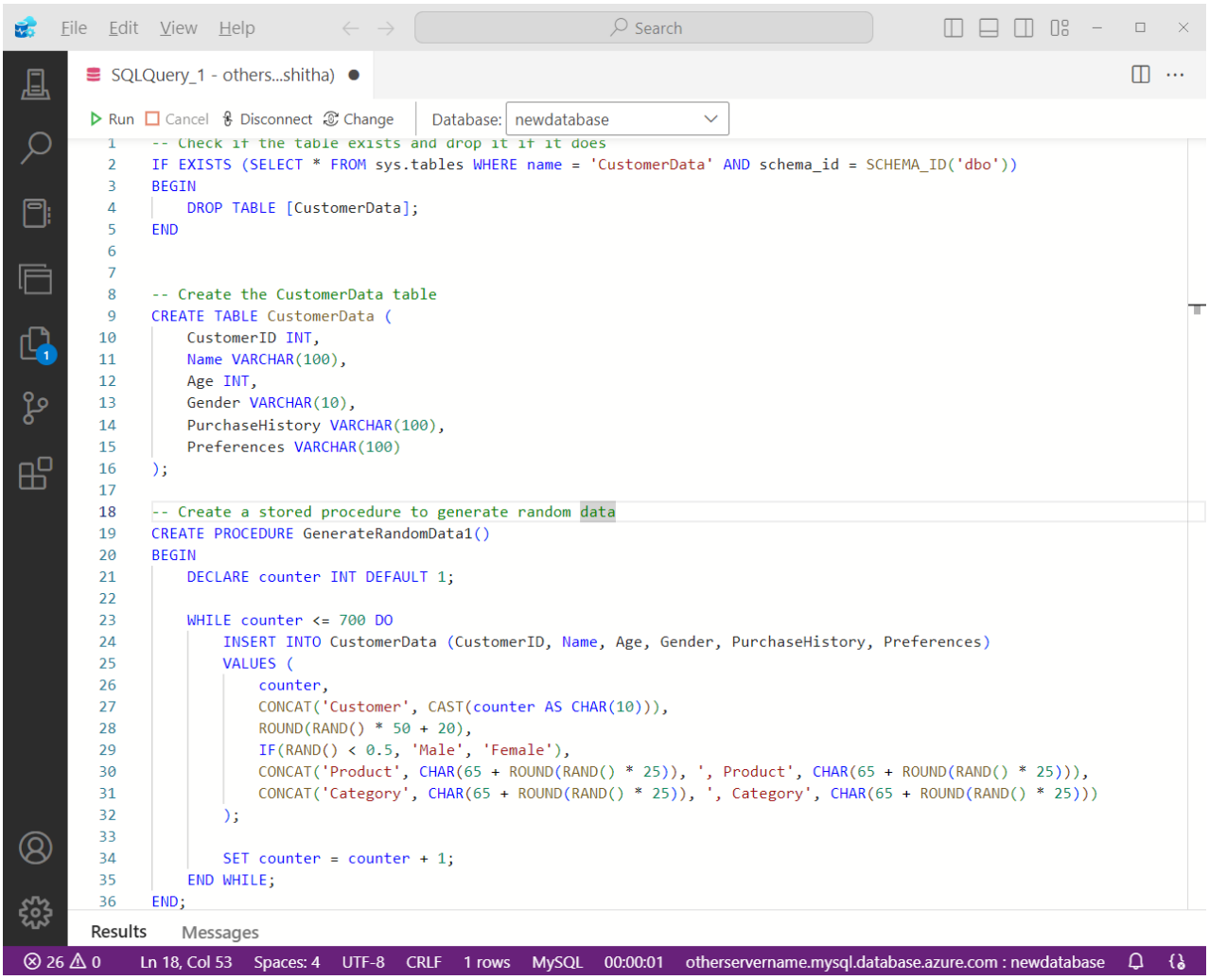

File Edit View Help
Search
CONNECTIONS
SERVERS
otherservername....
CustomerData_sourcedata.sql - dis
Run Cancel Disconnect Chan
Messages
AZURE
Ln 33, Col 1 Spaces: 4 UTF-8 CRLF 0 rows Executing query...
Connection
Recent
Browse
Clear List
otherservername.mysql.database.azure.com, newdatabase (rakshitha)
Connection Details
Connection type
MySQL
Server name *
otherservername.mysql.database.azure.com
Authentication type
Password
User name *
rakshitha
Password
Remember password
Database name
newdatabase
Server group
<Default>
Name (optional)
Advanced...
Connect
Cancel


File Edit View Help
Search
EXTENSIONS: MARKETPLACE
mysql
MySQL
1090ms
MySQL extension for Azure Data Studio
Microsoft
CustomerData_sourcedata.sql - others...shitha)
Extension: MySQL
MySQL v1.0.0
Microsoft
MySQL extension for Azure Data Studio
Disable
Uninstall
This extension is enabled globally.
DETAILS
FEATURE CONTRIBUTIONS
DEPENDENCIES
RUNTIME STATUS
MySQL extension for Azure Data Studio (preview)
Connect to, query, and manage MySQL databases with Azure Data Studio, a modern data editor available for Linux, MacOS, and Windows. This extension (preview) enables you to interact with MySQL using Azure Data Studio features like:
Connect to MySQL anywhere (on-premises, or VMs, on managed MySQL in other clouds or on Azure Database for MySQL - Flexible Server)
Searchable object explorer view for database objects, with auto-completion
Query authoring and editing with Intellisense, syntax highlighting and code snippets
View query results and save to csv, JSON, xml, or Excel
Integrated terminal for Bash, PowerShell, and cmd.exe
Source control integration with Git
Customize dashboards and insight widgets using SQL
Server groups for organizing connections
Customizable keyboard shortcuts, multi-tab support, color theme options
See our quickstart for a step-by-step guide to get started with MySQL using Azure Data Studio.
Extension Resources
Repository
License
More Info
Last released 2023-6-22, 00:00:00
Last updated 2024-6-17, 13:59:56
Identifier microsoft.azuredatas mysql
Choose SQL Language

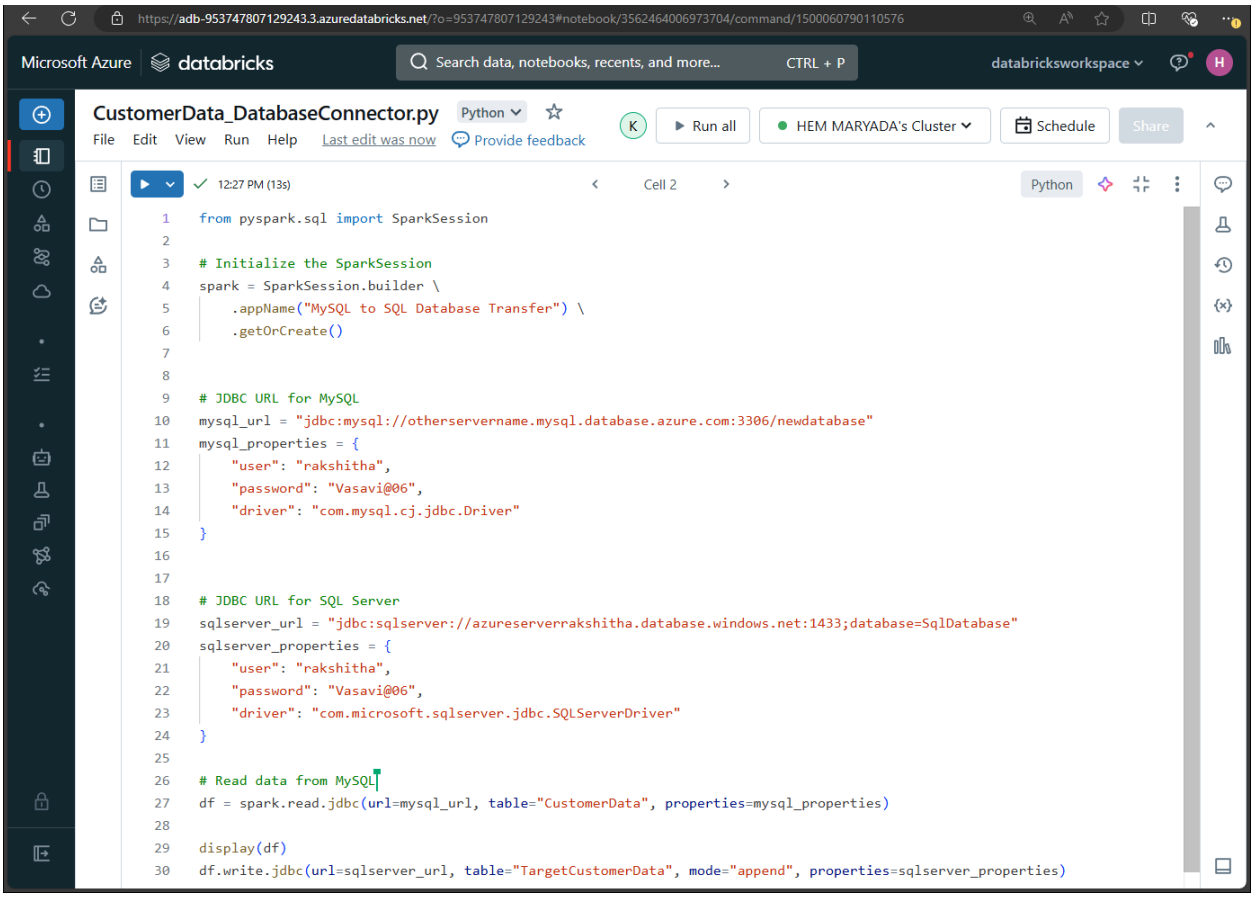

File Edit View Help
CONNECTIONS
SERVERS
otherservername.mysql.database.azure.com, newdatabase (rakshitha)
Character Sets
Databases
System Databases
Users
Tablespaces

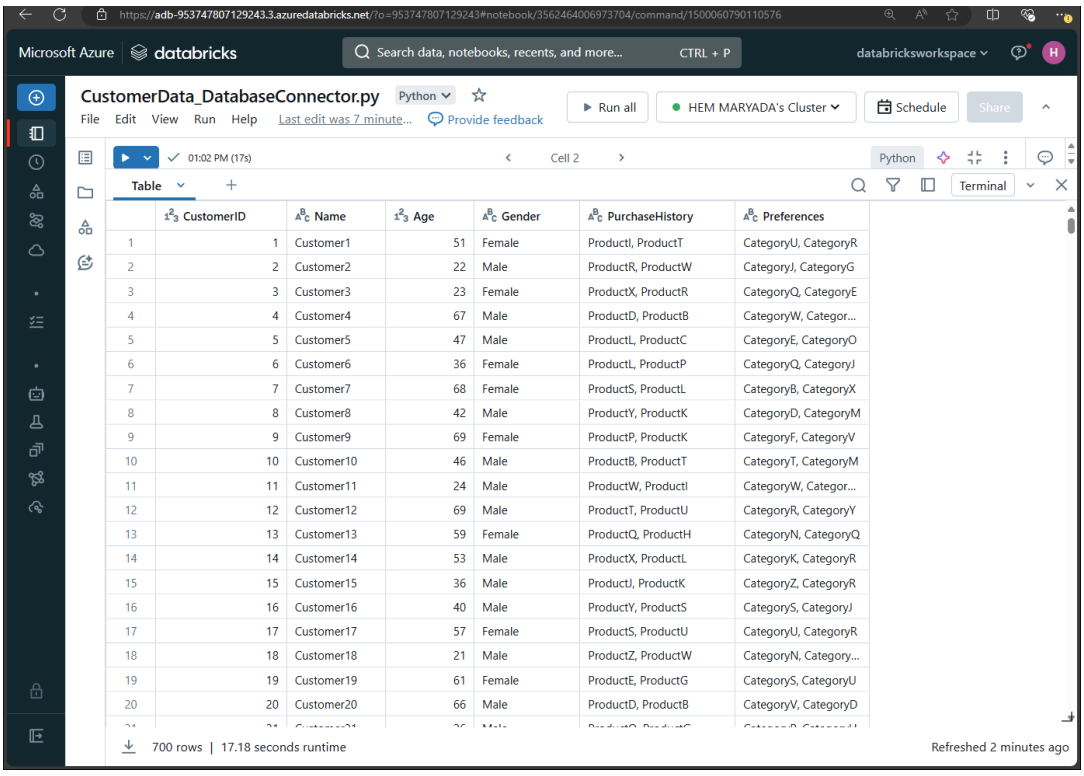

File Edit View Help
Search
CONNECTIONS
SERVERS
otherservername.mysql.database.azure.com, newdatabase (rakshitha)
Character Sets
Databases
newdatabase
Tables
customerdata
Columns
my_row_id
CustomerID
Name
Age
Gender
PurchaseHistory
Preferences
Constraints
Indexes
Triggers
Views
Stored Procedures
Functions
Events
System Databases
Users
Tablespaces
AZURE
SQLQuery_1 - others...shitha)
Run Cancel Disconnect Change Database:
newdatabase
41
42
43 -- Output the generated customerdata
44 SELECT count(*) FROM CustomerData;
45
46
Results Messages
count(*)
1 700
26 0 Ln 42, Col 1 Spaces: 4 UTF-8 CRLF 1 rows MySQL 00:00:01 otherservername.mysql.database.azure.com : newdatabase

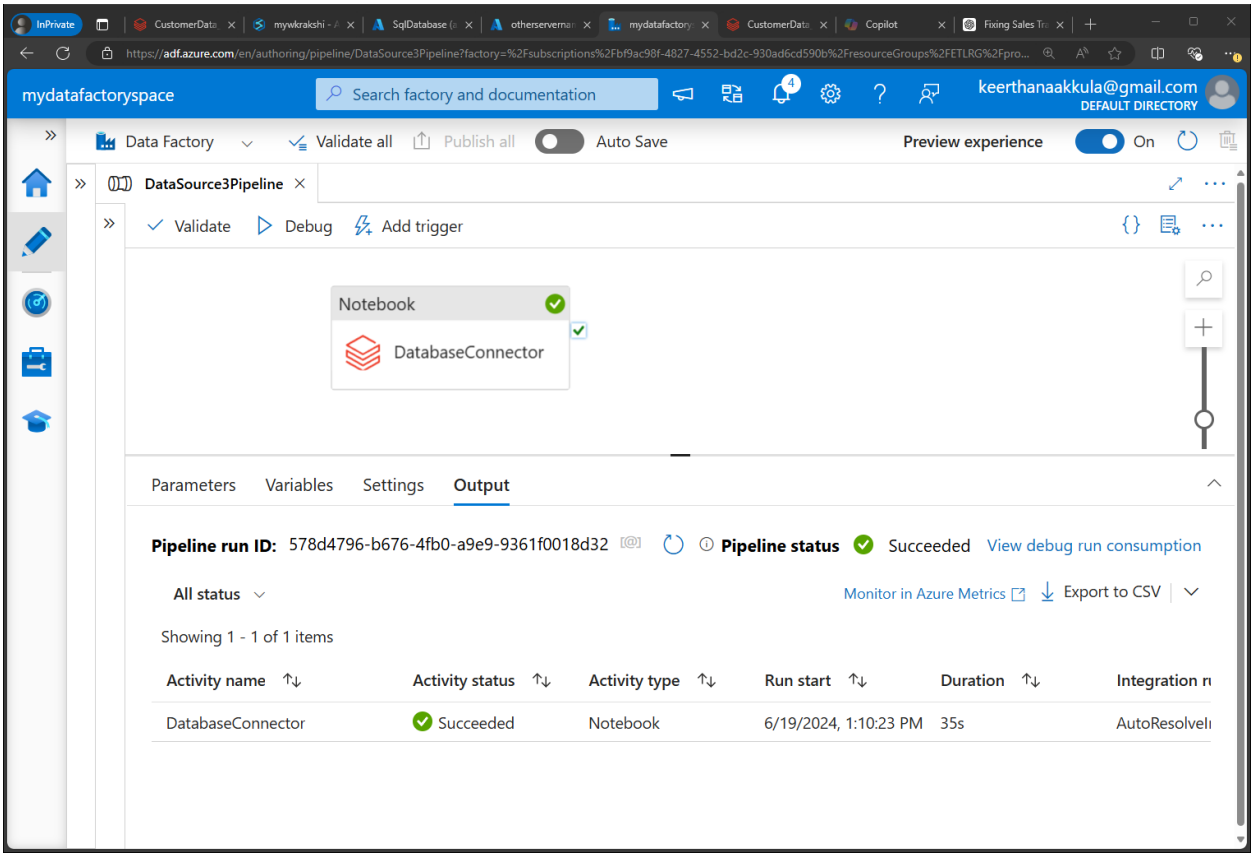
Created a random customer data in Azure Database for MySQL flexible server via Azure data studio

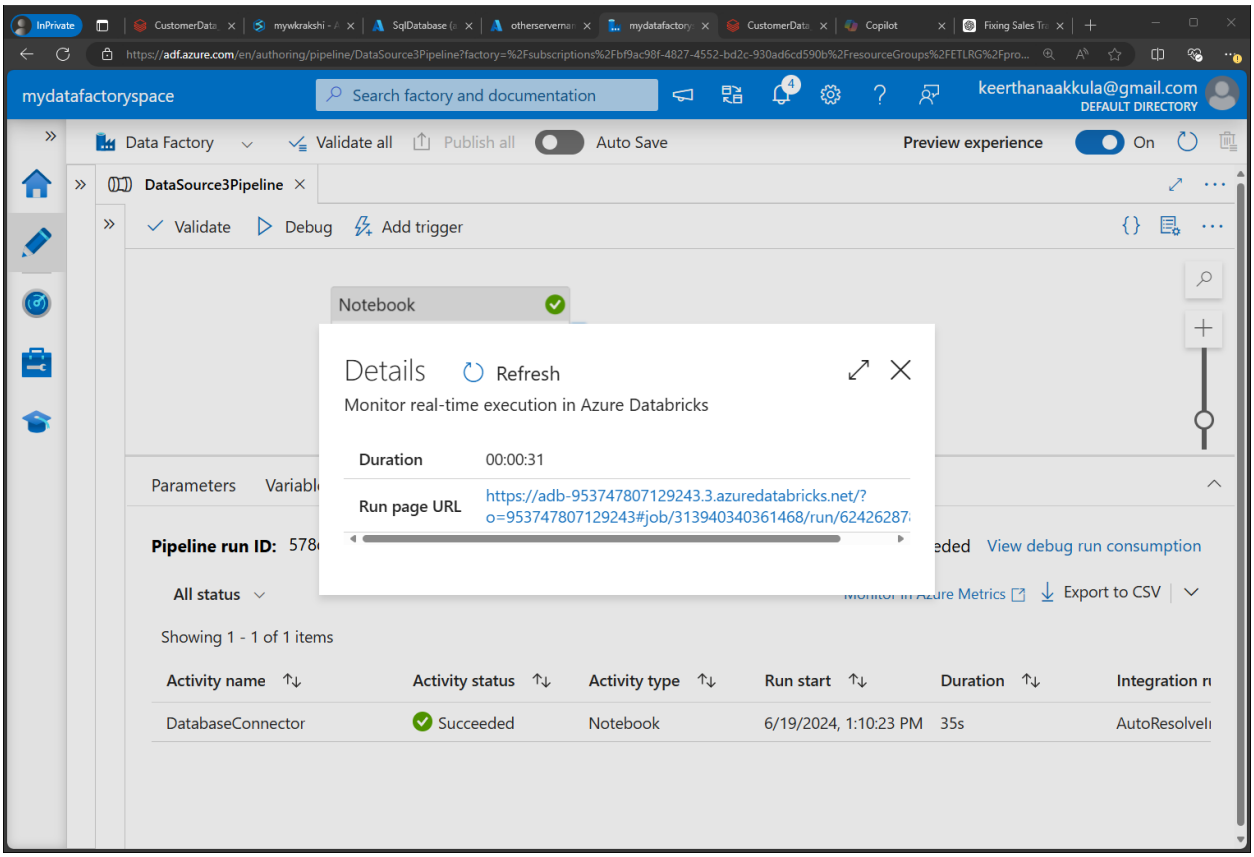



CustomerData_DatabaseConnector.py.html

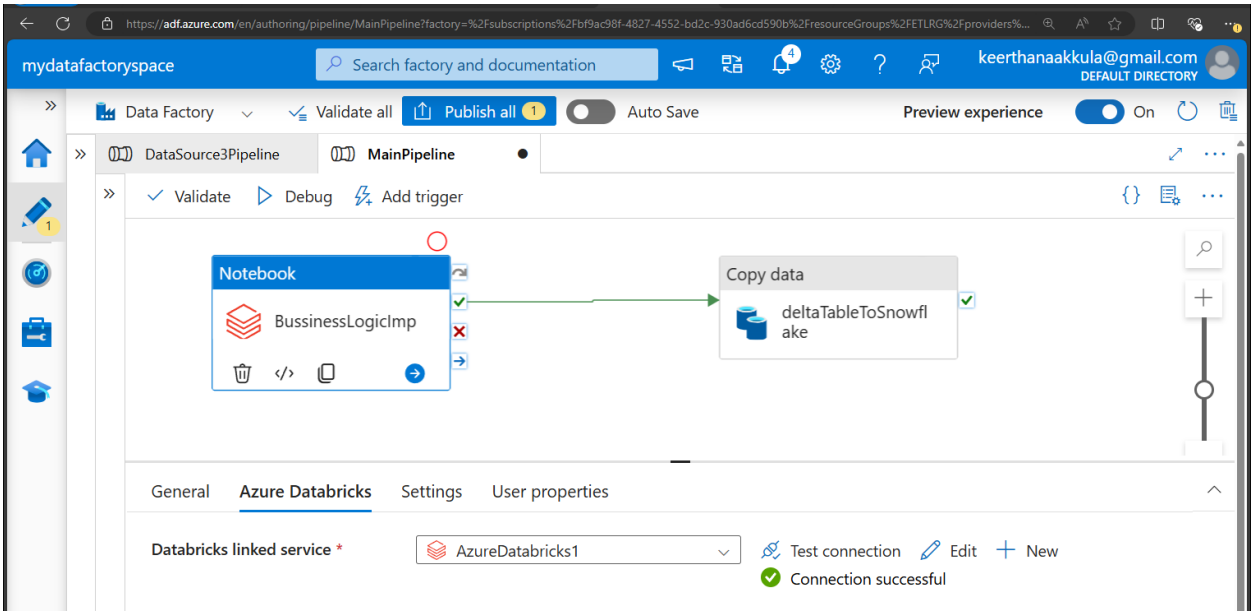

https://adb-953747807129243.3.azuredatabricks.net/?o=953747807129243#notebook/3562464006973704/command/1500060790110576
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
databricksworkspace
CustomerData_DatabaseConnector.py
Python
File Edit View Run Help
Last edit was now
Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
12:27 PM (13s)
Cell 2
Python
from pyspark.sql import SparkSession
# Initialize the SparkSession
spark = SparkSession.builder \
.appName("MySQL to SQL Database Transfer") \
.getOrCreate()
# JDBC URL for MySQL
mysql_url = "jdbc:mysql://otherservername.mysql.database.azure.com:3306/newdatabase"
mysql_properties = {
"user": "rakshitha",
"password": "Vasavi@06",
"driver": "com.mysql.cj.jdbc.Driver"
}
# JDBC URL for SQL Server
sqlserver_url = "jdbc:sqlserver://azureserverrakshitha.database.windows.net:1433;database=SqlDatabase"
sqlserver_properties = {
"user": "rakshitha",
"password": "Vasavi@06",
"driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
}
# Read data from MySQL
df = spark.read.jdbc(url=mysql_url, table="CustomerData", properties=mysql_properties)
display(df)
df.write.jdbc(url=sqlserver_url, table="TargetCustomerData", mode="append", properties=sqlserver_properties)


https://adb-953747807129243.3.azuredatabricks.net/?o=953747807129243#notebook/3562464006973704/command/1500060790110576
Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
databricksworkspace
CustomerData_DatabaseConnector.py
Python
File Edit View Run Help
Last edit was 7 minute...
Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
01:02 PM (17s)
Cell 2
Python
Table
Terminal
CustomerID Name Age Gender PurchaseHistory Preferences
1 1 Customer1 51 Female ProductI, ProductT CategoryU, CategoryR
2 2 Customer2 22 Male ProductR, ProductW CategoryJ, CategoryG
3 3 Customer3 23 Female ProductX, ProductR CategoryQ, CategoryE
4 4 Customer4 67 Male ProductD, ProductB CategoryW, Categor...
5 5 Customer5 47 Male ProductL, ProductC CategoryE, CategoryO
6 6 Customer6 36 Female ProductL, ProductP CategoryQ, CategoryJ
7 7 Customer7 68 Female ProductS, ProductL CategoryB, CategoryX
8 8 Customer8 42 Male ProductY, ProductK CategoryD, CategoryM
9 9 Customer9 69 Female ProductP, ProductK CategoryF, CategoryV
10 10 Customer10 46 Male ProductB, ProductT CategoryT, CategoryM
11 11 Customer11 24 Male ProductW, ProductI CategoryW, Categor...
12 12 Customer12 69 Male ProductT, ProductU CategoryR, CategoryY
13 13 Customer13 59 Female ProductQ, ProductH CategoryN, CategoryQ
14 14 Customer14 53 Male ProductX, ProductL CategoryK, CategoryR
15 15 Customer15 36 Male ProductJ, ProductK CategoryZ, CategoryR
16 16 Customer16 40 Male ProductY, ProductS CategoryS, CategoryJ
17 17 Customer17 57 Female ProductS, ProductU CategoryU, CategoryR
18 18 Customer18 21 Male ProductZ, ProductW CategoryN, Category...
19 19 Customer19 61 Female ProductE, ProductG CategoryS, CategoryU
20 20 Customer20 66 Male ProductD, ProductB CategoryV, CategoryD
700 rows | 17.18 seconds runtime
Refreshed 2 minutes ago


mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
Validate
Debug
Add trigger
Notebook
DatabaseConnector
Parameters
Variables
Settings
Output
Pipeline run ID: 578d4796-b676-4fb0-a9e9-9361f0018d32
Pipeline status
Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Integration r
DatabaseConnector
Succeeded
Notebook
6/19/2024, 1:10:23 PM
35s
AutoResolveI

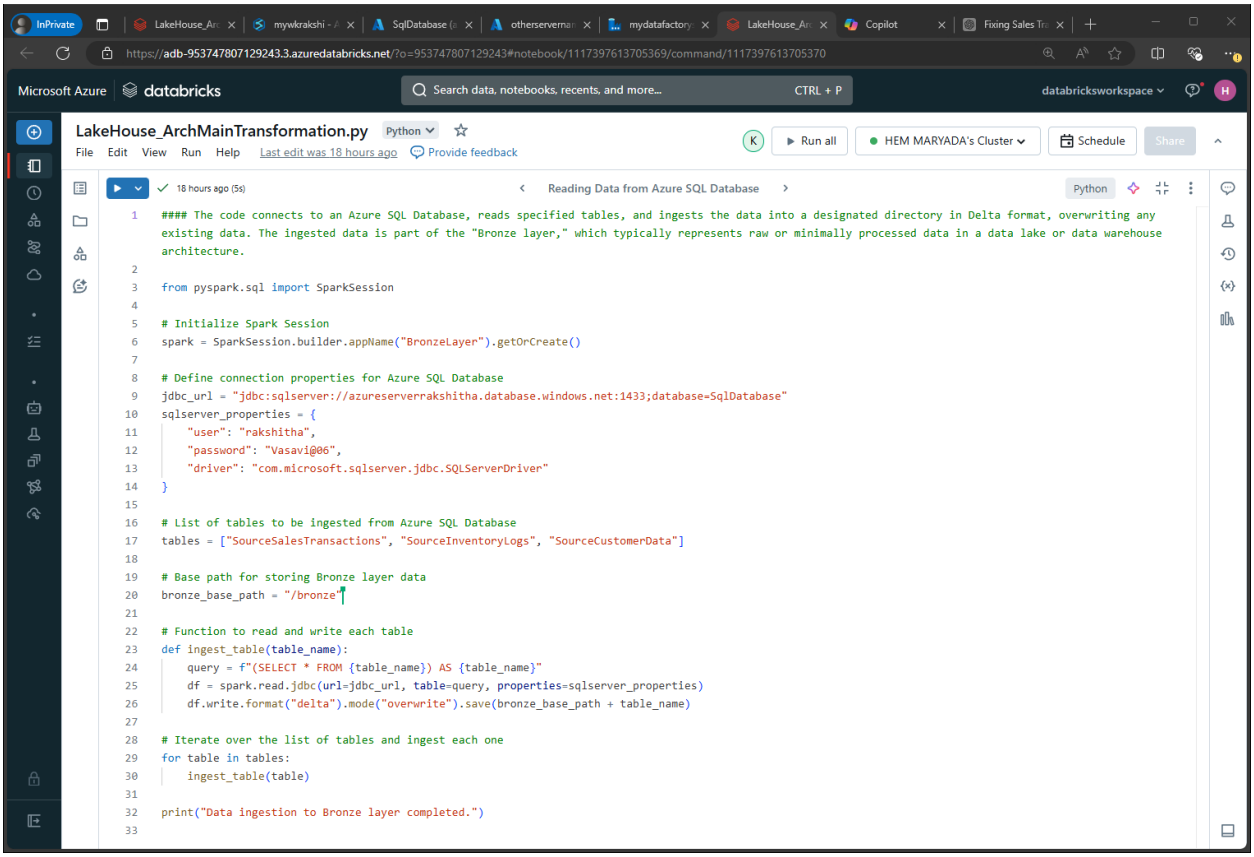

mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Auto Save
Preview experience
On
DataSource3Pipeline
Validate
Debug
Add trigger
Notebook
Details
Refresh
Monitor real-time execution in Azure Databricks
Duration
00:00:31
Run page URL
https://adb-953747807129243.3.azuredatabricks.net/?o=953747807129243#job/313940340361468/run/62426287
Parameters
Pipeline run ID:
View debug run consumption
All status
Export to CSV
Showing 1 - 1 of 1 items
Activity name
Activity status
Activity type
Run start
Duration
Integration r
DatabaseConnector
Succeeded
Notebook
6/19/2024, 1:10:23 PM
35s
AutoResolveI

2. **Data Storage: Now all the data sources are stored in Azure SQL Database:** Now all the 3 data sources are in Azure SQL Database, I have created a pipeline to copy the data from Azure SQL Database to Snowflake
3. **Data Transformation: Pyspark (notebook):** Created a notebook that implements the **Lakehouse Architecture** (medallion)
4. **Data Warehousing and Analytics: Snowflake, Power BI**, I have created a pipeline to copy the data from Azure SQL Database to Snowflake



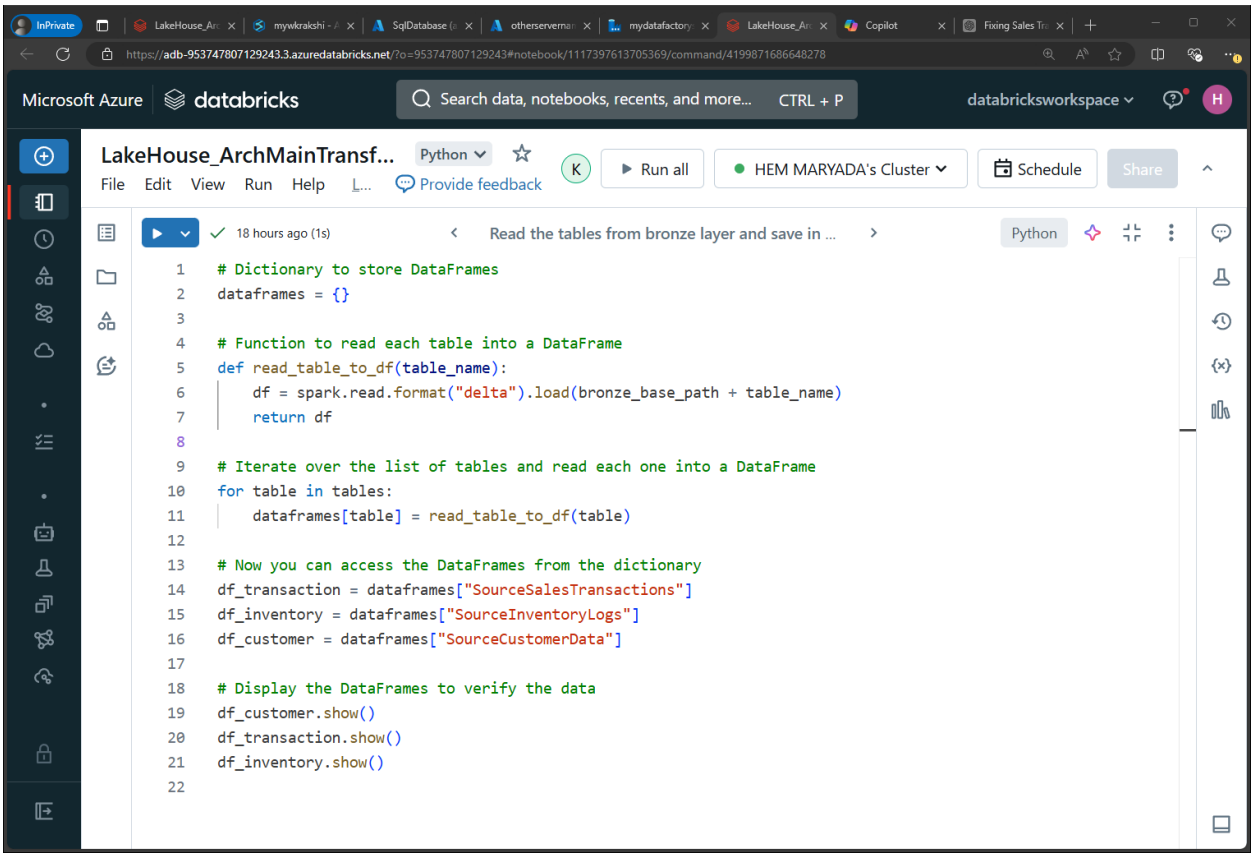

The Medallion Architecture, also known as the Lakehouse Architecture, is a data management pattern that organizes data in layers to improve data quality, processing efficiency, and maintainability. This architecture typically involves three layers:

1. **Bronze (Raw Data) Layer**: Stores raw data in its original format.
2. **Silver (Refined Data) Layer**: Contains cleaned and processed data, ready for analytics.
3. **Gold (Serving) Layer**: Consists of highly refined data used for reporting and dashboards.



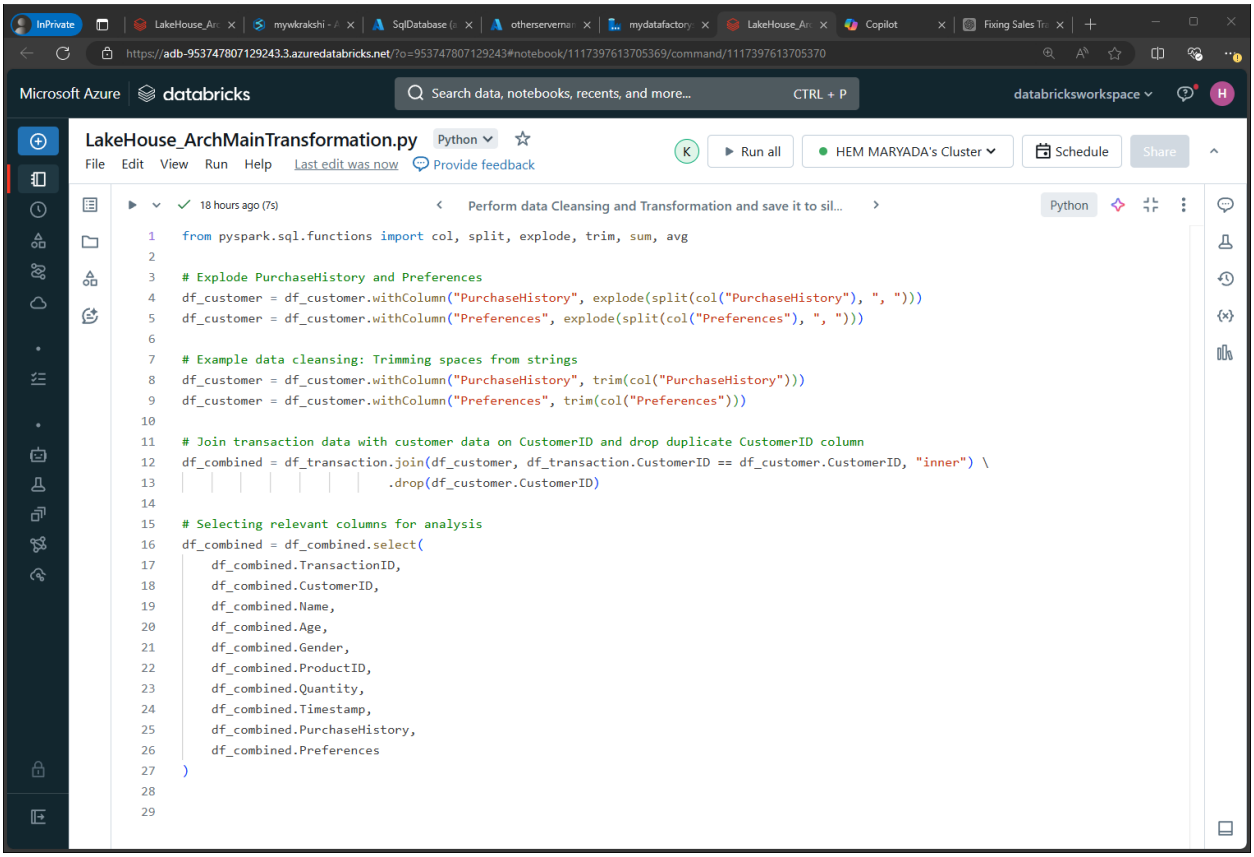
LakeHouse_ArchMainTransformation.py

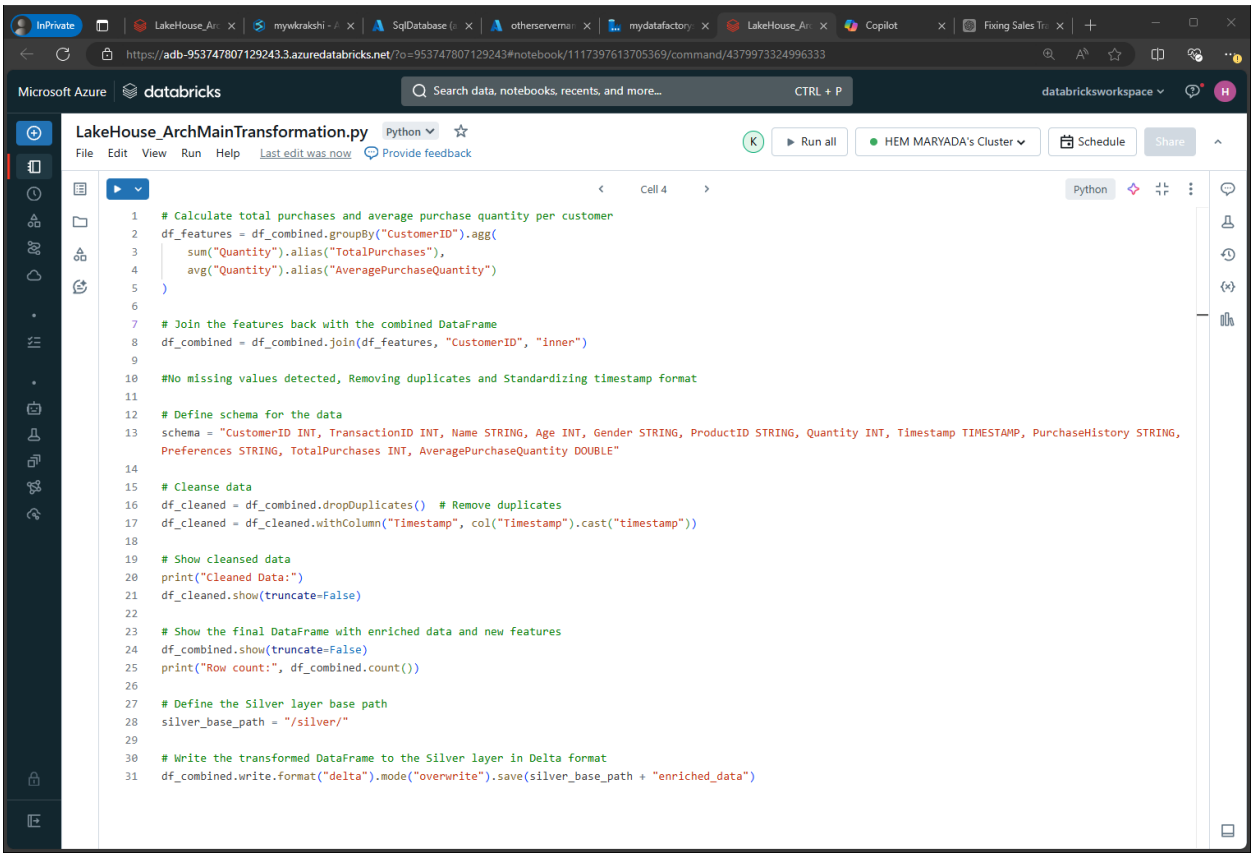
```python
#### The code connects to an Azure SQL Database, reads specified tables, and ingests the data into a designated directory in Delta format, overwriting any existing data. The ingested data is part of the "Bronze layer," which typically represents raw or minimally processed data in a data lake or data warehouse architecture.

from pyspark.sql import SparkSession

# Initialize Spark Session
spark = SparkSession.builder.appName("BronzeLayer").getOrCreate()

# Define connection properties for Azure SQL Database
jdbc_url = "jdbc:sqlserver://azureserverrakshitha.database.windows.net:1433;database=SqlDatabase"
sqlserver_properties = {
    "user": "rakshitha",
    "password": "Vasavi@06",
    "driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
}

# List of tables to be ingested from Azure SQL Database
tables = ["SourceSalesTransactions", "SourceInventoryLogs", "SourceCustomerData"]

# Base path for storing Bronze layer data
bronze_base_path = "/bronze"

# Function to read and write each table
def ingest_table(table_name):
    query = f"(SELECT * FROM {table_name}) AS {table_name}"
    df = spark.read.jdbc(url=jdbc_url, table=query, properties=sqlserver_properties)
    df.write.format("delta").mode("overwrite").save(bronze_base_path + table_name)

# Iterate over the list of tables and ingest each one
for table in tables:
    ingest_table(table)

print("Data ingestion to Bronze layer completed.")
```

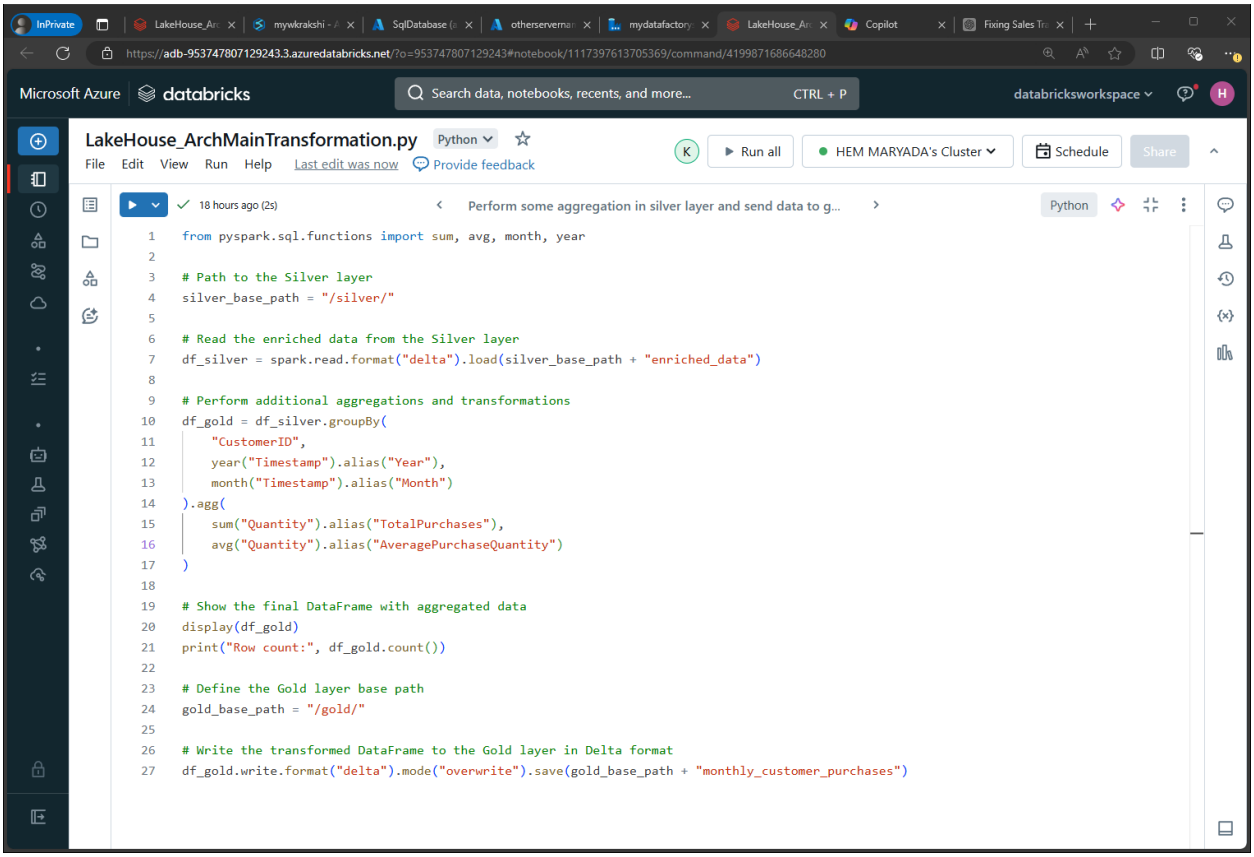


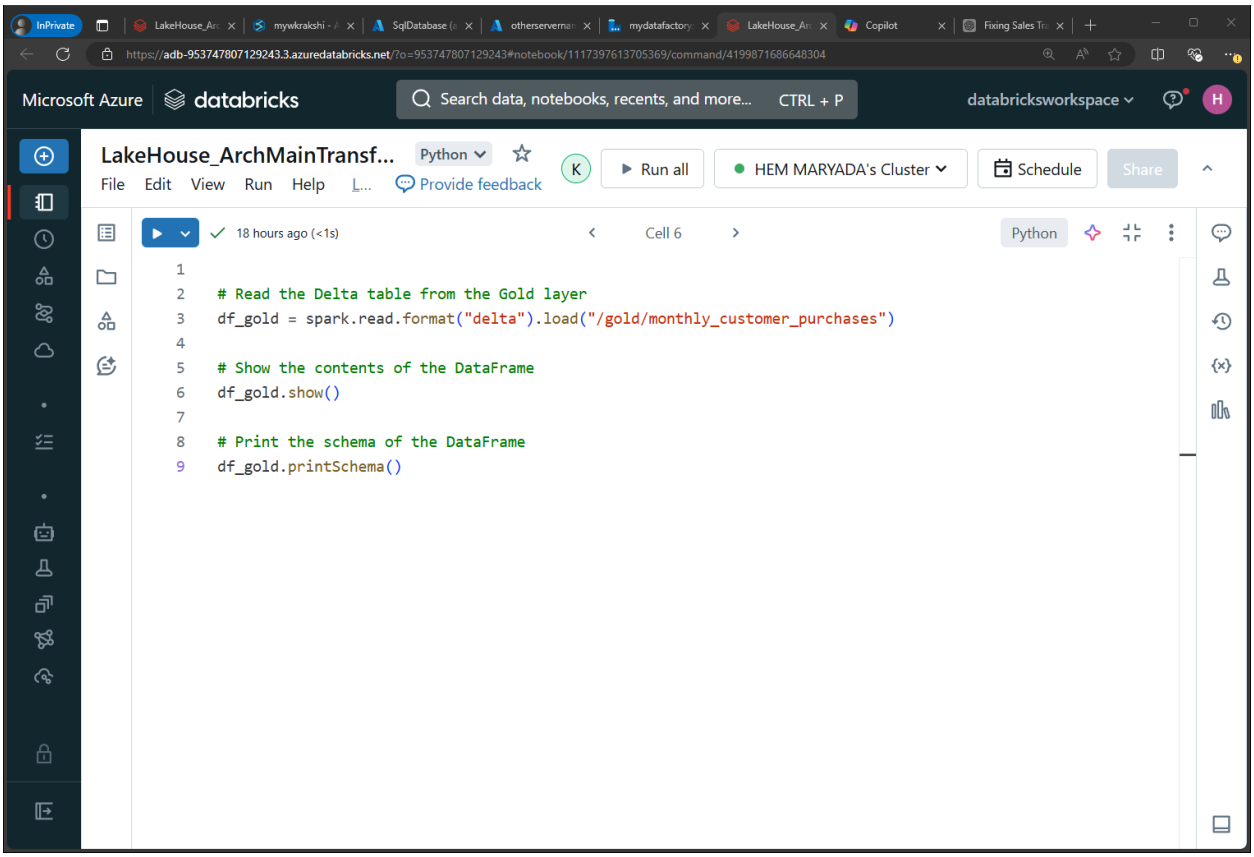
LakeHouse_ArchMainTransf...

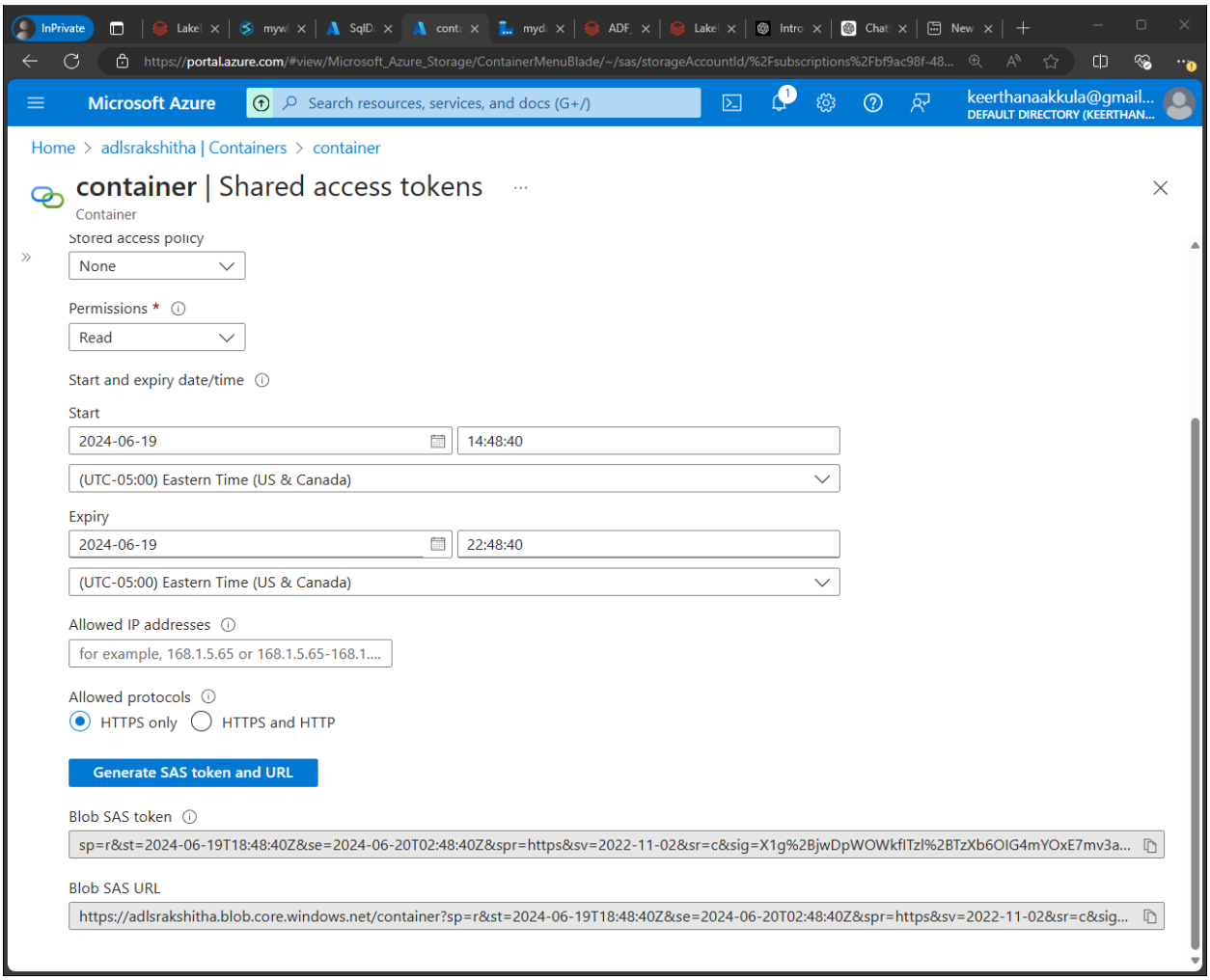
```python
# Dictionary to store DataFrames
dataframes = {}

# Function to read each table into a DataFrame
def read_table_to_df(table_name):
    df = spark.read.format("delta").load(bronze_base_path + table_name)
    return df

# Iterate over the list of tables and read each one into a DataFrame
for table in tables:
    dataframes[table] = read_table_to_df(table)

# Now you can access the DataFrames from the dictionary
df_transaction = dataframes["SourceSalesTransactions"]
df_inventory = dataframes["SourceInventoryLogs"]
df_customer = dataframes["SourceCustomerData"]

# Display the DataFrames to verify the data
df_customer.show()
df_transaction.show()
df_inventory.show()
```


LakeHouse_ArchMainTransformation.py
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
18 hours ago (7s)
Perform data Cleansing and Transformation and save it to sil...
from pyspark.sql.functions import col, split, explode, trim, sum, avg
# Explode PurchaseHistory and Preferences
df_customer = df_customer.withColumn("PurchaseHistory", explode(split(col("PurchaseHistory"), ", ")))
df_customer = df_customer.withColumn("Preferences", explode(split(col("Preferences"), ", ")))
# Example data cleansing: Trimming spaces from strings
df_customer = df_customer.withColumn("PurchaseHistory", trim(col("PurchaseHistory")))
df_customer = df_customer.withColumn("Preferences", trim(col("Preferences")))
# Join transaction data with customer data on CustomerID and drop duplicate CustomerID column
df_combined = df_transaction.join(df_customer, df_transaction.CustomerID == df_customer.CustomerID, "inner") \
.drop(df_customer.CustomerID)
# Selecting relevant columns for analysis
df_combined = df_combined.select(
df_combined.TransactionID,
df_combined.CustomerID,
df_combined.Name,
df_combined.Age,
df_combined.Gender,
df_combined.ProductID,
df_combined.Quantity,
df_combined.Timestamp,
df_combined.PurchaseHistory,
df_combined.Preferences
)

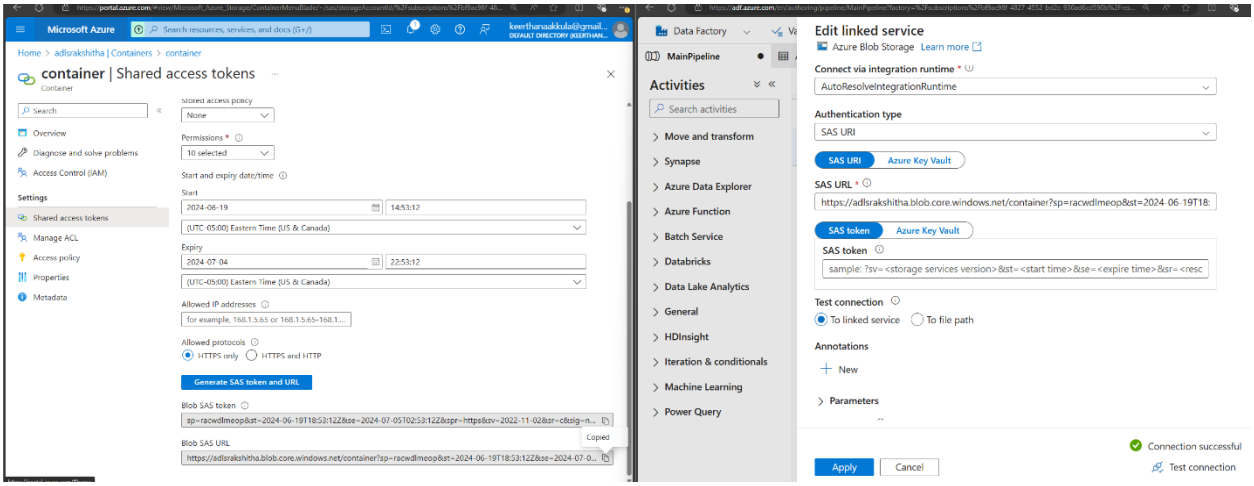

LakeHouse_ArchMainTransformation.py
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
Cell 4
# Calculate total purchases and average purchase quantity per customer
df_features = df_combined.groupBy("CustomerID").agg(
sum("Quantity").alias("TotalPurchases"),
avg("Quantity").alias("AveragePurchaseQuantity")
)
# Join the features back with the combined DataFrame
df_combined = df_combined.join(df_features, "CustomerID", "inner")
#No missing values detected, Removing duplicates and Standardizing timestamp format
# Define schema for the data
schema = "CustomerID INT, TransactionID INT, Name STRING, Age INT, Gender STRING, ProductID STRING, Quantity INT, Timestamp TIMESTAMP, PurchaseHistory STRING, Preferences STRING, TotalPurchases INT, AveragePurchaseQuantity DOUBLE"
# Cleanse data
df_cleaned = df_combined.dropDuplicates() # Remove duplicates
df_cleaned = df_cleaned.withColumn("Timestamp", col("Timestamp").cast("timestamp"))
# Show cleansed data
print("Cleaned Data:")
df_cleaned.show(truncate=False)
# Show the final DataFrame with enriched data and new features
df_combined.show(truncate=False)
print("Row count:", df_combined.count())
# Define the Silver layer base path
silver_base_path = "/silver/"
# Write the transformed DataFrame to the Silver layer in Delta format
df_combined.write.format("delta").mode("overwrite").save(silver_base_path + "enriched_data")

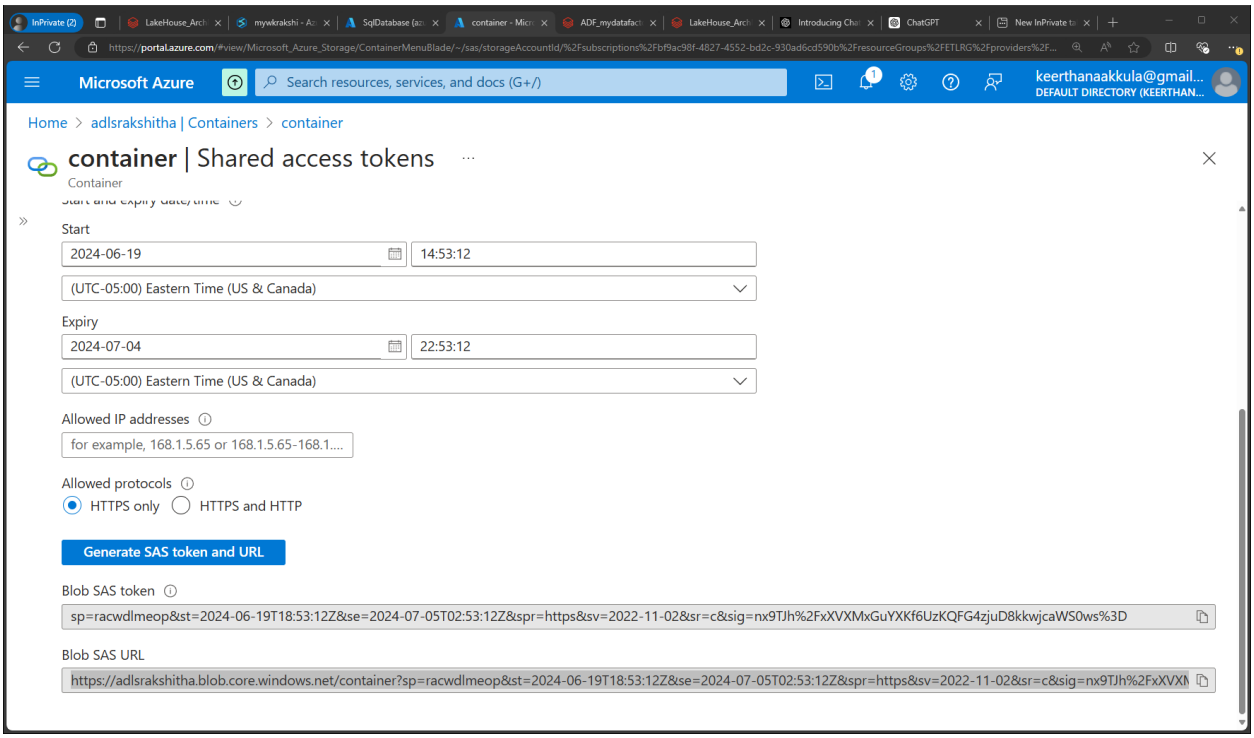

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
databricksworkspace
LakeHouse_ArchMainTransformation.py
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
18 hours ago (2s)
Perform some aggregation in silver layer and send data to g...
Python
from pyspark.sql.functions import sum, avg, month, year
# Path to the Silver layer
silver_base_path = "/silver/"
# Read the enriched data from the Silver layer
df_silver = spark.read.format("delta").load(silver_base_path + "enriched_data")
# Perform additional aggregations and transformations
df_gold = df_silver.groupBy(
"CustomerID",
year("Timestamp").alias("Year"),
month("Timestamp").alias("Month")
).agg(
sum("Quantity").alias("TotalPurchases"),
avg("Quantity").alias("AveragePurchaseQuantity")
)
# Show the final DataFrame with aggregated data
display(df_gold)
print("Row count:", df_gold.count())
# Define the Gold layer base path
gold_base_path = "/gold/"
# Write the transformed DataFrame to the Gold layer in Delta format
df_gold.write.format("delta").mode("overwrite").save(gold_base_path + "monthly_customer_purchases")

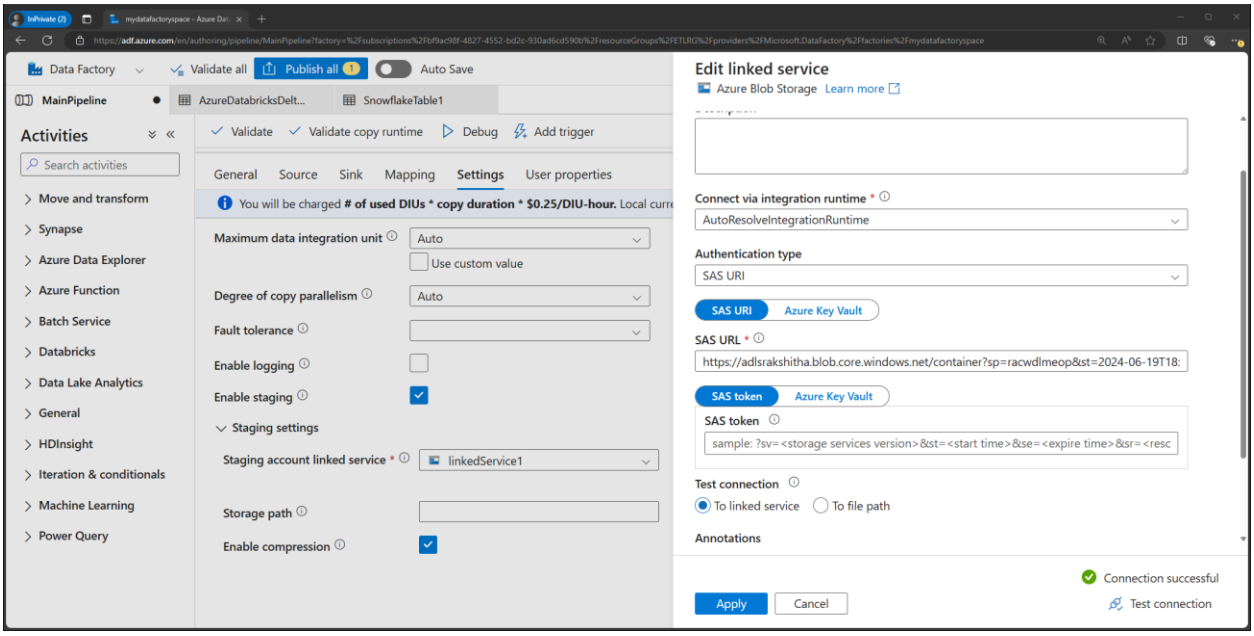

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
databricksworkspace
LakeHouse_ArchMainTransf...
Python
File Edit View Run Help L... Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
18 hours ago (<1s)
Cell 6
Python
# Read the Delta table from the Gold layer
df_gold = spark.read.format("delta").load("/gold/monthly_customer_purchases")
# Show the contents of the DataFrame
df_gold.show()
# Print the schema of the DataFrame
df_gold.printSchema()

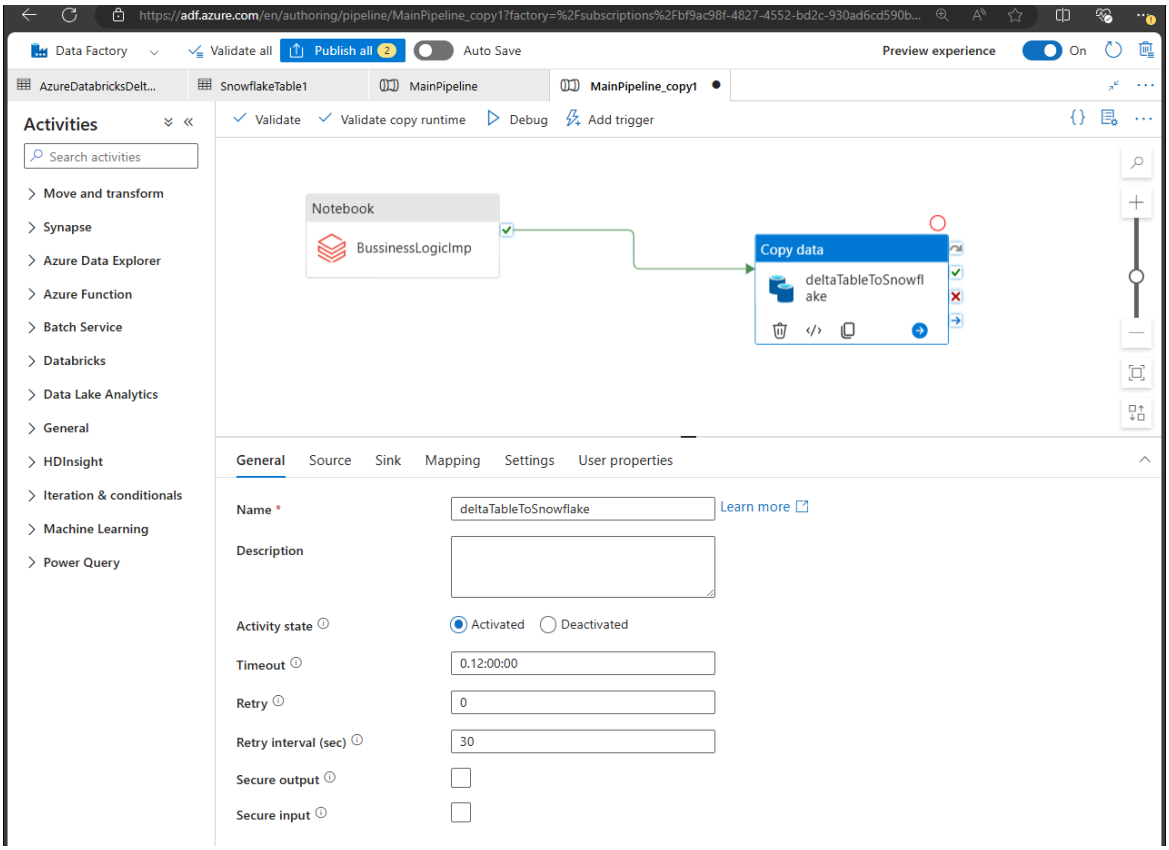


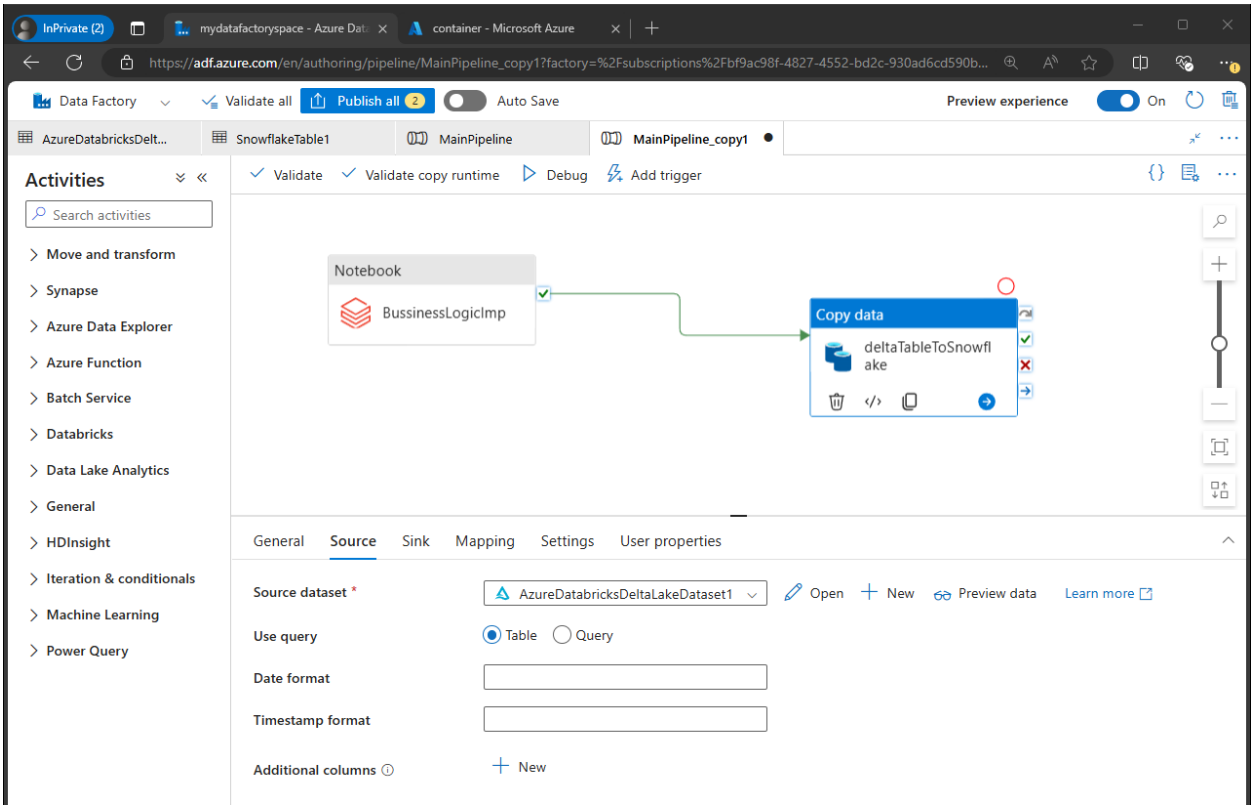

Keep the Blob SAS URL in the SAS URL and leave the SAS token blank in data factory as shown below

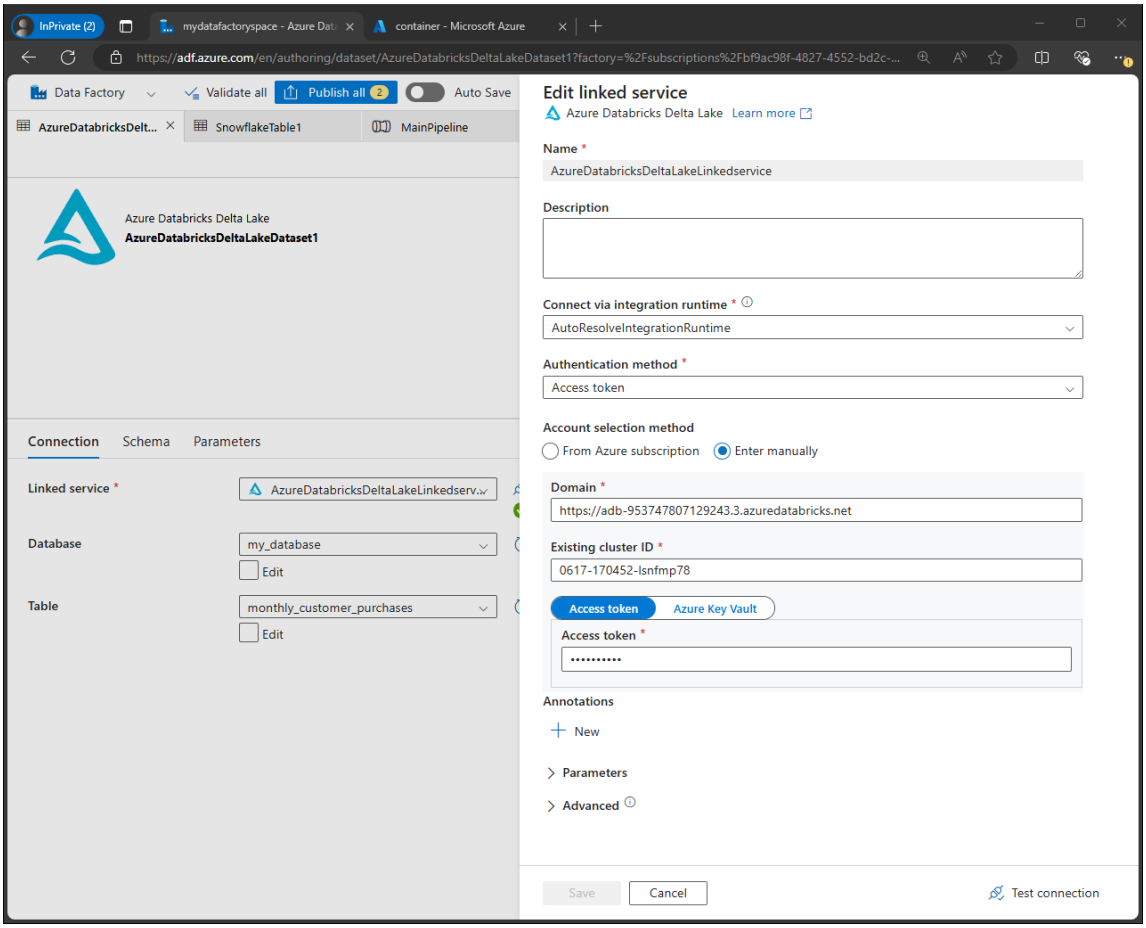


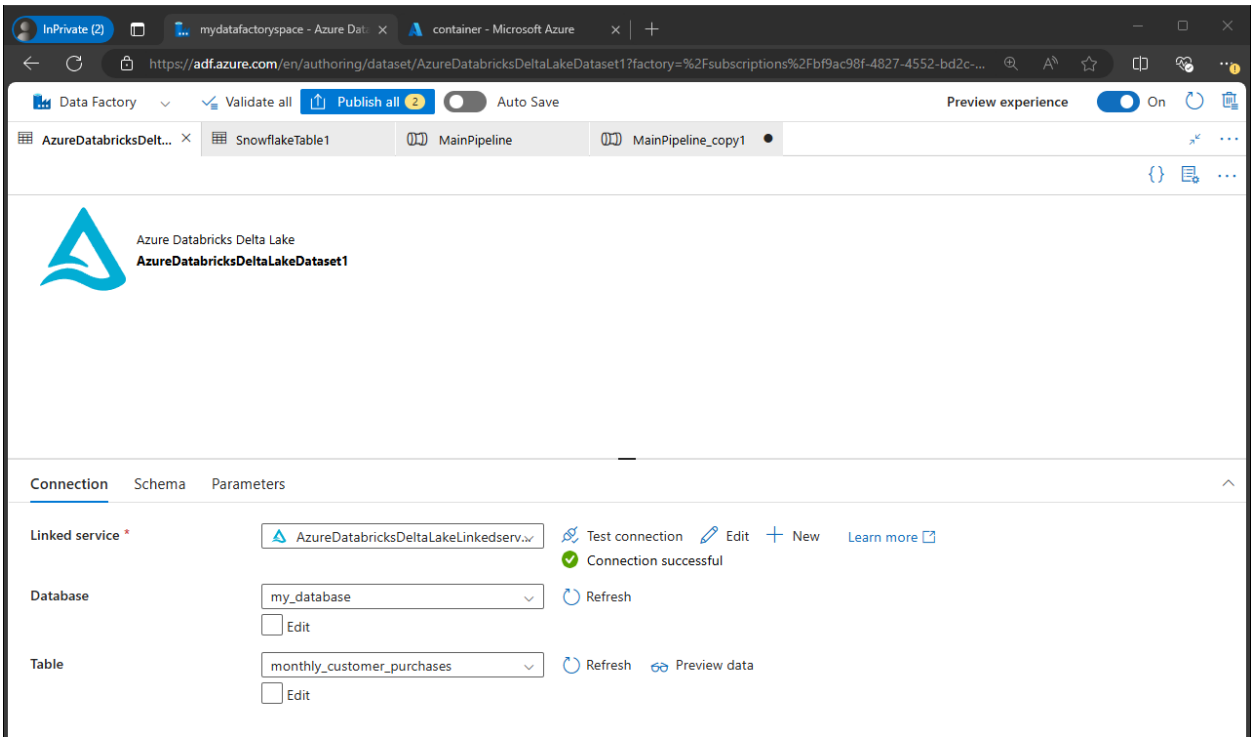

Microsoft Azure
Search resources, services, and docs (G+/)
keerthanaakkula@gmail...
DEFAULT DIRECTORY (KEERTHAN...
Home > adlsrakshitha | Containers > container
container | Shared access tokens
Container
Start
2024-06-19
14:53:12
(UTC-05:00) Eastern Time (US & Canada)
Expiry
2024-07-04
22:53:12
(UTC-05:00) Eastern Time (US & Canada)
Allowed IP addresses
for example, 168.1.5.65 or 168.1.5.65-168.1....
Allowed protocols
HTTPS only
HTTPS and HTTP
Generate SAS token and URL
Blob SAS token
sp=racwdlmeop&st=2024-06-19T18:53:12Z&se=2024-07-05T02:53:12Z&spr=https&sv=2022-11-02&sr=c&sig=nx9TJh%2FxXVXMxGuYXKf6UzKQFG4zjuD8kkwjcaWS0ws%3D
Blob SAS URL
https://adlsrakshitha.blob.core.windows.net/container?sp=racwdlmeop&st=2024-06-19T18:53:12Z&se=2024-07-05T02:53:12Z&spr=https&sv=2022-11-02&sr=c&sig=nx9TJh%2FxXVX

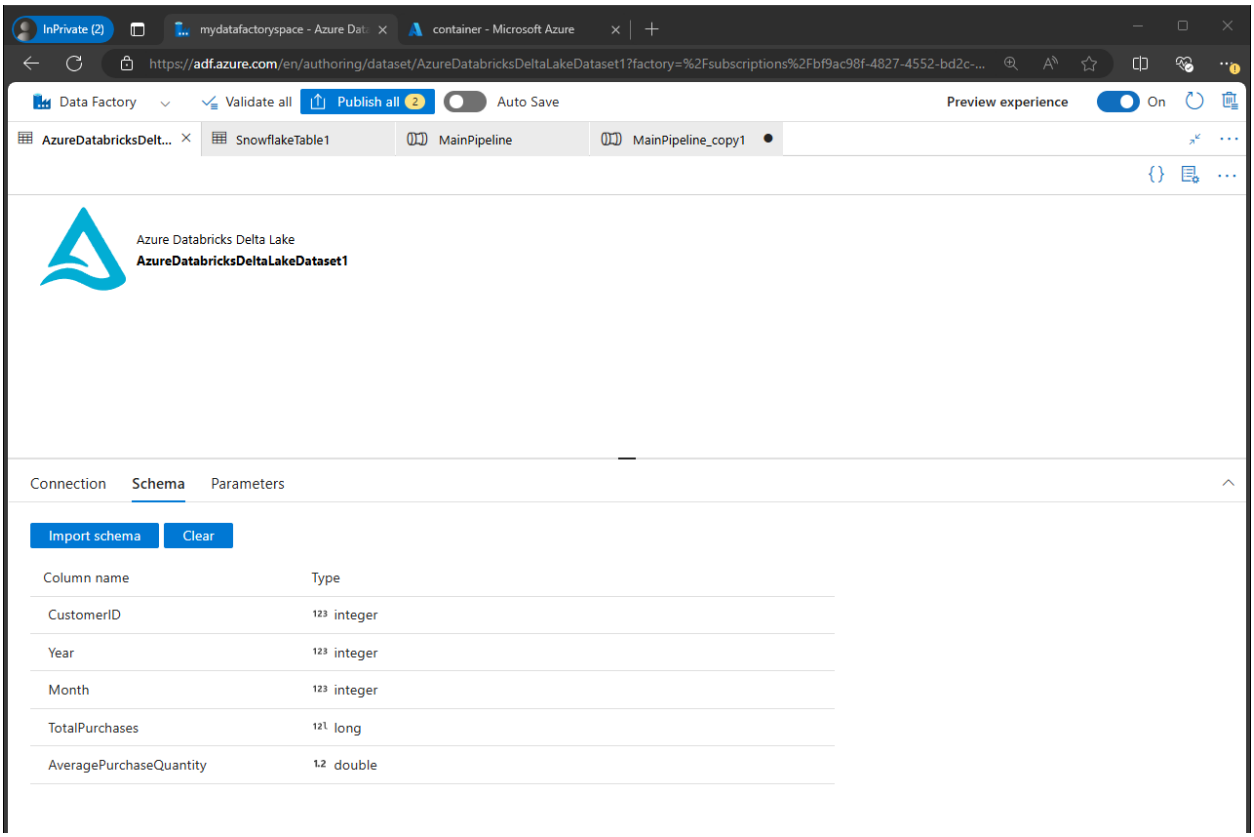

Data Factory
Validate all
Publish all
Auto Save
MainPipeline
AzureDatabricksDelt...
SnowflakeTable1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
General
Source
Sink
Mapping
Settings
User properties
Maximum data integration unit
Auto
Use custom value
Degree of copy parallelism
Auto
Fault tolerance
Enable logging
Enable staging
Staging settings
Staging account linked service
linkedService1
Storage path
Enable compression
Edit linked service
Azure Blob Storage
Learn more
Connect via integration runtime
AutoResolveIntegrationRuntime
Authentication type
SAS URI
SAS URI
Azure Key Vault
SAS URL
https://adlsrakshitha.blob.core.windows.net/container?sp=racwdlmeop&st=2024-06-19T18:
SAS token
Azure Key Vault
SAS token
sample: ?sv=<storage services version>&st=<start time>&se=<expire time>&sr=<resc
Test connection
To linked service
To file path
Annotations
Connection successful
Test connection
Apply
Cancel

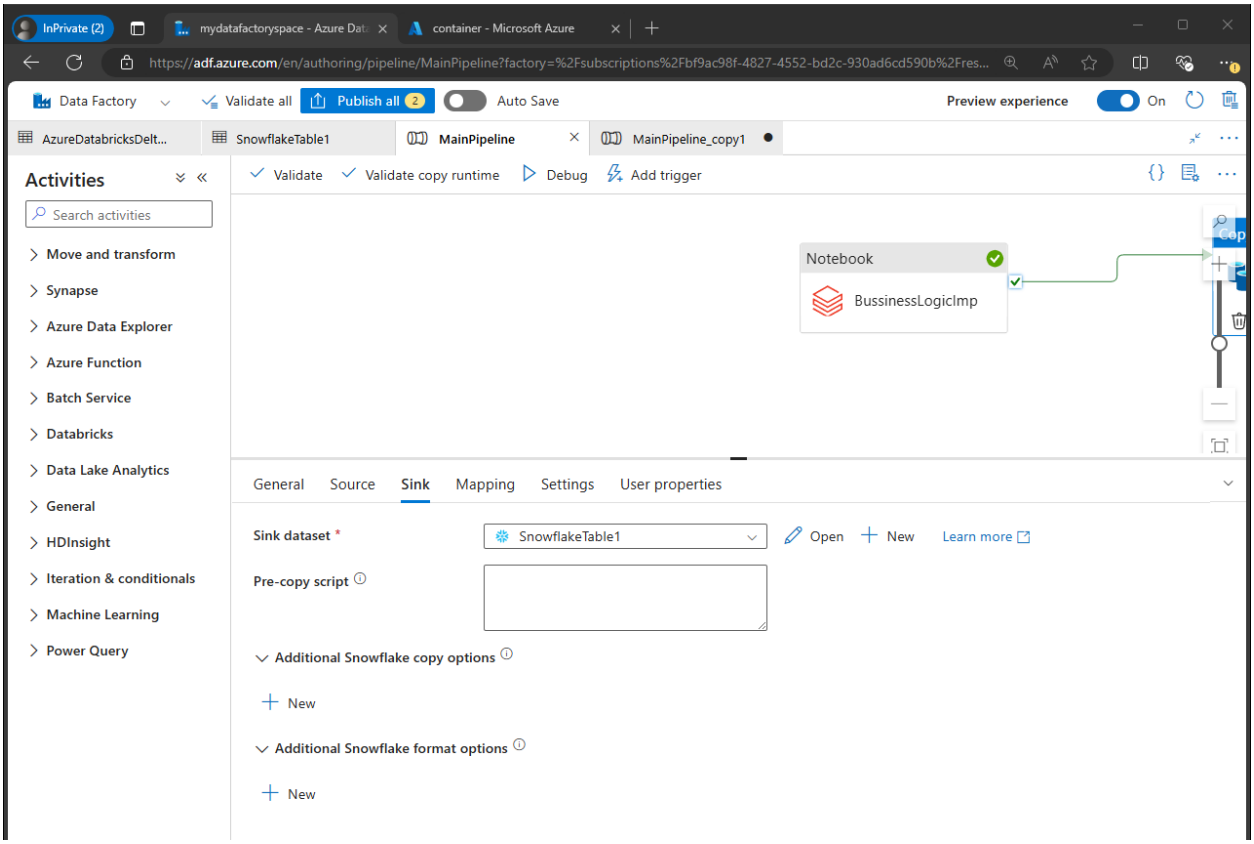

https://adf.azure.com/en/authoring/pipeline/MainPipeline_copy1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Notebook
BussinessLogicImp
Copy data
deltaTableToSnowflake
General
Source
Sink
Mapping
Settings
User properties
Name *
deltaTableToSnowflake
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.12:00:00
Retry
0
Retry interval (sec)
30
Secure output
Secure input

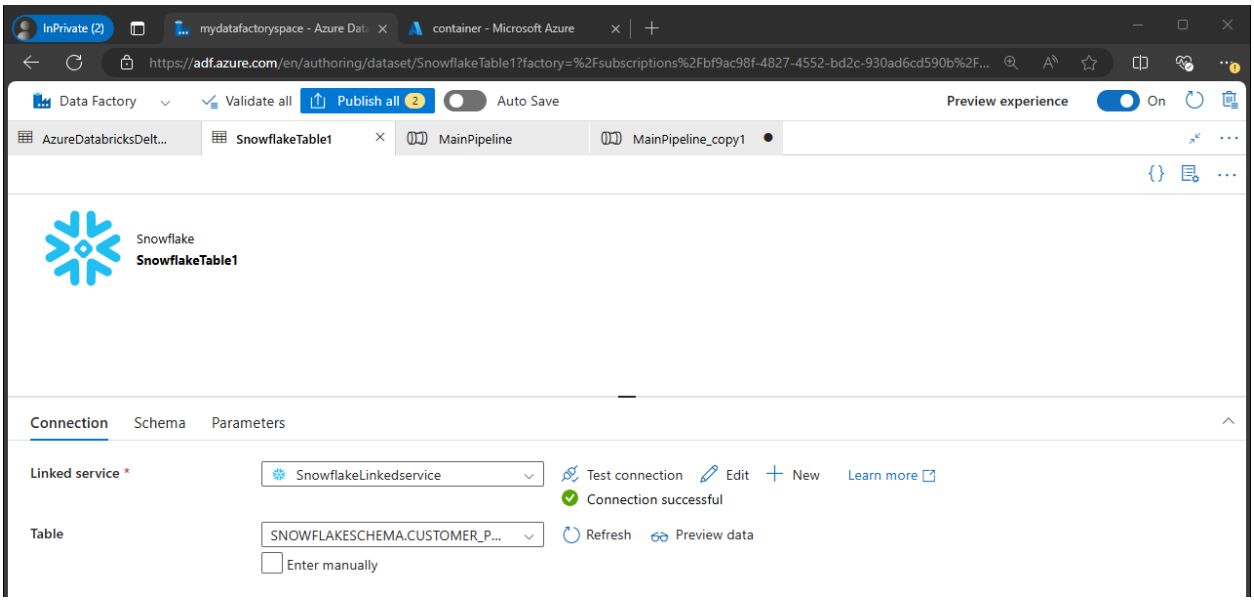

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/pipeline/MainPipeline_copy1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Notebook
BussinessLogicImp
Copy data
deltaTableToSnowflake
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
AzureDatabricksDeltaLakeDataset1
Open
New
Preview data
Learn more
Use query
Table
Query
Date format
Timestamp format
Additional columns
New

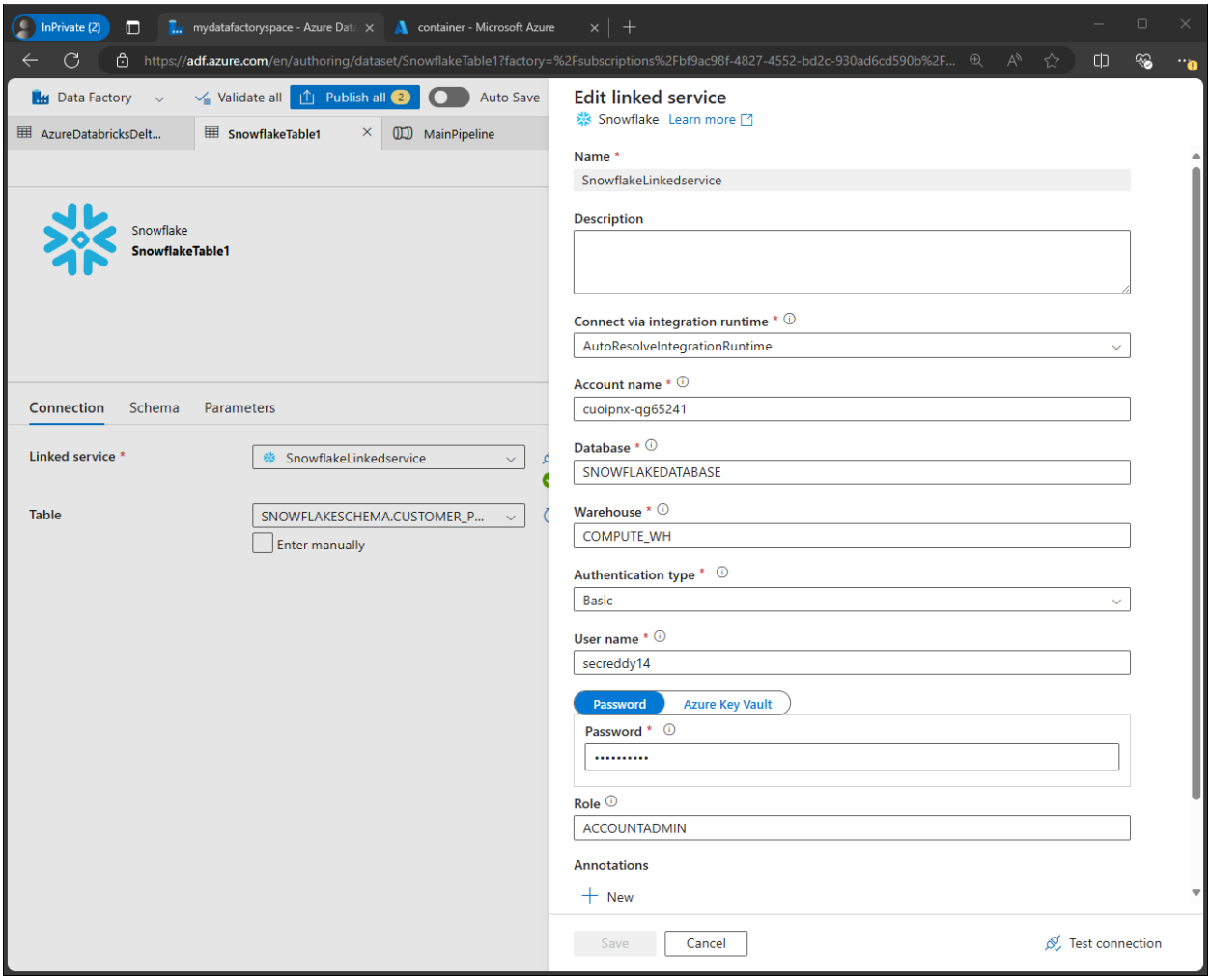

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/AzureDatabricksDeltaLakeDataset1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-...
Data Factory
Validate all
Publish all 2
Auto Save
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
Azure Databricks Delta Lake
AzureDatabricksDeltaLakeDataset1
Connection
Schema
Parameters
Linked service *
AzureDatabricksDeltaLakeLinkedserv...
Database
my_database
Edit
Table
monthly_customer_purchases
Edit
Edit linked service
Azure Databricks Delta Lake
Learn more
Name *
AzureDatabricksDeltaLakeLinkedservice
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Authentication method *
Access token
Account selection method
From Azure subscription
Enter manually
Domain *
https://adb-953747807129243.3.azuredatabricks.net
Existing cluster ID *
0617-170452-lsnfmp78
Access token
Azure Key Vault
Access token *
Annotations
New
Parameters
Advanced
Save
Cancel
Test connection

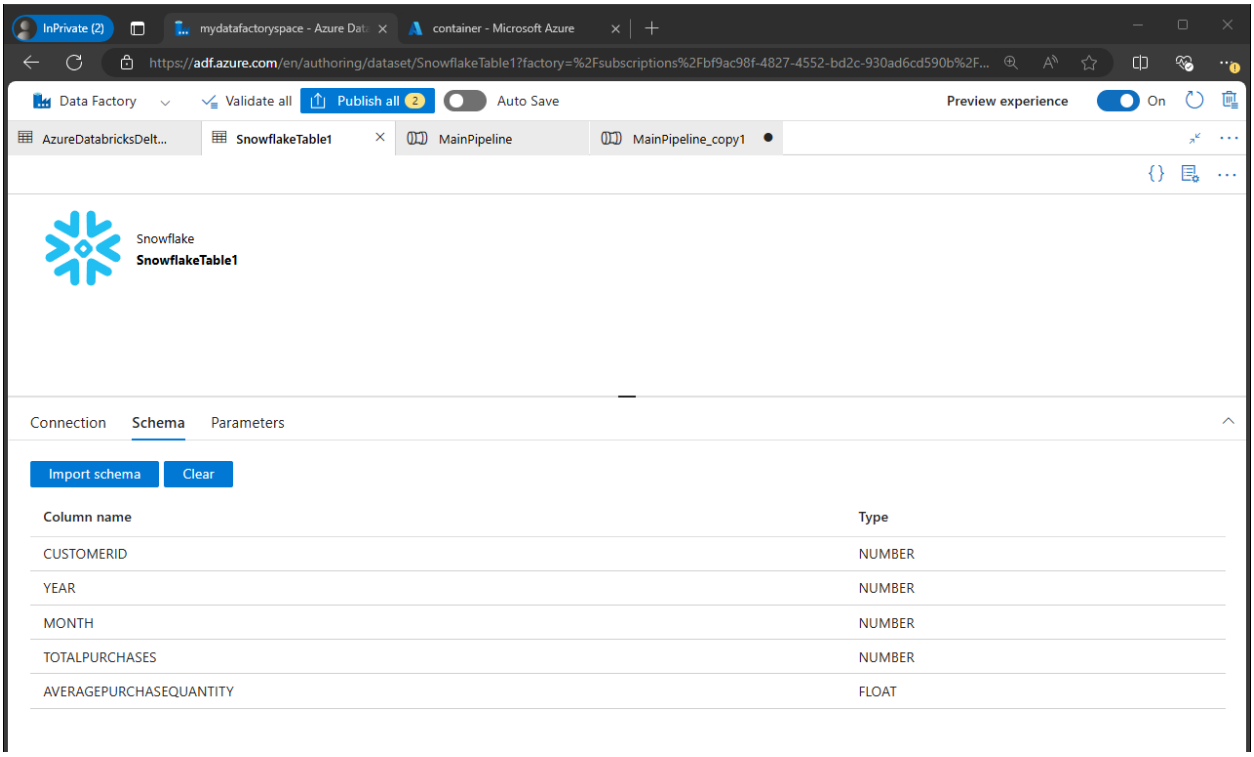

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/AzureDatabricksDeltaLakeDataset1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Azure Databricks Delta Lake
AzureDatabricksDeltaLakeDataset1
Connection
Schema
Parameters
Linked service *
AzureDatabricksDeltaLakeLinkedserv...
Test connection
Edit
New
Learn more
Connection successful
Database
my_database
Refresh
Edit
Table
monthly_customer_purchases
Refresh
Preview data
Edit

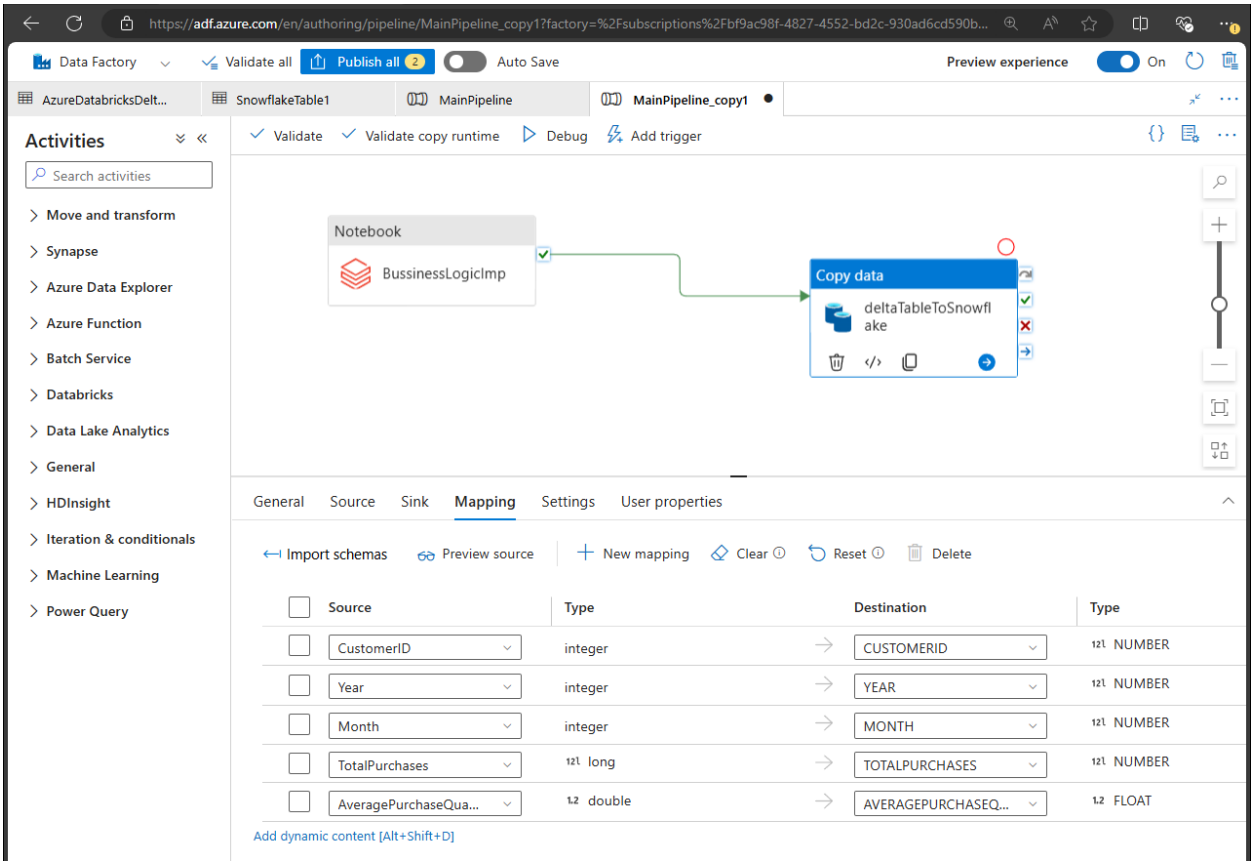

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/AzureDatabricksDeltaLakeDataset1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Azure Databricks Delta Lake
AzureDatabricksDeltaLakeDataset1
Connection
Schema
Parameters
Import schema
Clear
Column name
Type
CustomerID
integer
Year
integer
Month
integer
TotalPurchases
long
AveragePurchaseQuantity
double


InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/pipeline/MainPipeline?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2Fres...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Notebook
BussinessLogicImp
General
Source
Sink
Mapping
Settings
User properties
Sink dataset *
SnowflakeTable1
Open
New
Learn more
Pre-copy script
Additional Snowflake copy options
New
Additional Snowflake format options
New


InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/SnowflakeTable1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2F...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Snowflake
SnowflakeTable1
Connection
Schema
Parameters
Linked service *
SnowflakeLinkedservice
Test connection
Edit
New
Learn more
Connection successful
Table
SNOWFLAKESCHEMA.CUSTOMER_P...
Refresh
Preview data
Enter manually

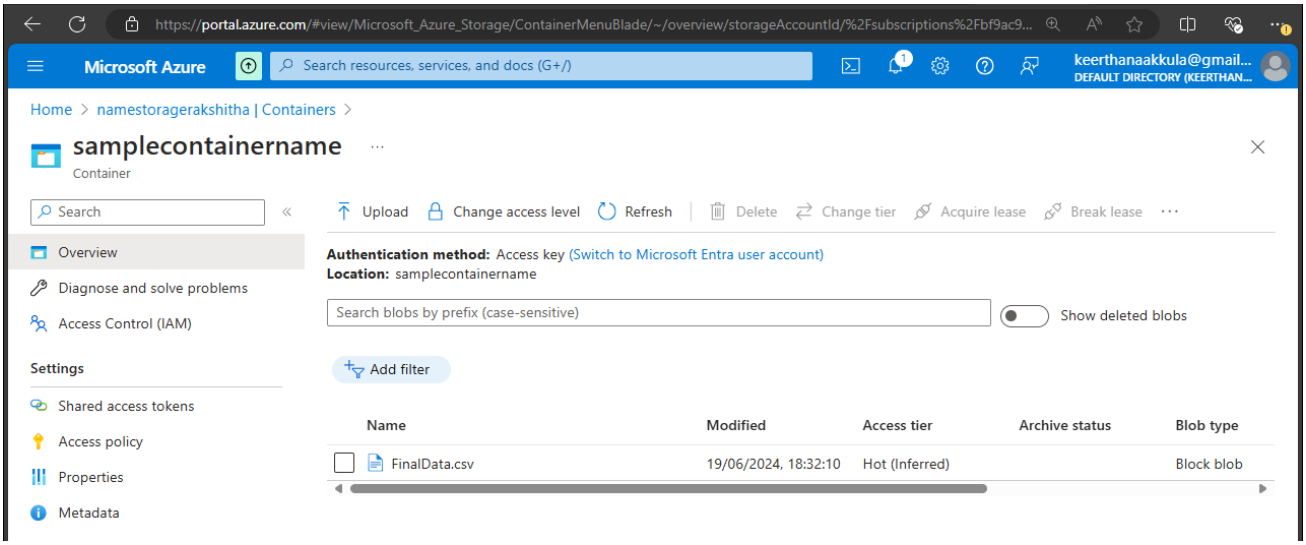

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/SnowflakeTable1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2F...
Data Factory
Validate all
Publish all 2
Auto Save
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
Snowflake
SnowflakeTable1
Connection
Schema
Parameters
Linked service *
SnowflakeLinkedservice
Table
SNOWFLAKESCHEMA.CUSTOMER_P...
Enter manually
Edit linked service
Snowflake
Learn more
Name *
SnowflakeLinkedservice
Description
Connect via integration runtime *
AutoResolveIntegrationRuntime
Account name *
cuoipnx-qg65241
Database *
SNOWFLAKEDATABASE
Warehouse *
COMPUTE_WH
Authentication type *
Basic
User name *
secreddy14
Password
Azure Key Vault
Password *
Role
ACCOUNTADMIN
Annotations
New
Save
Cancel
Test connection

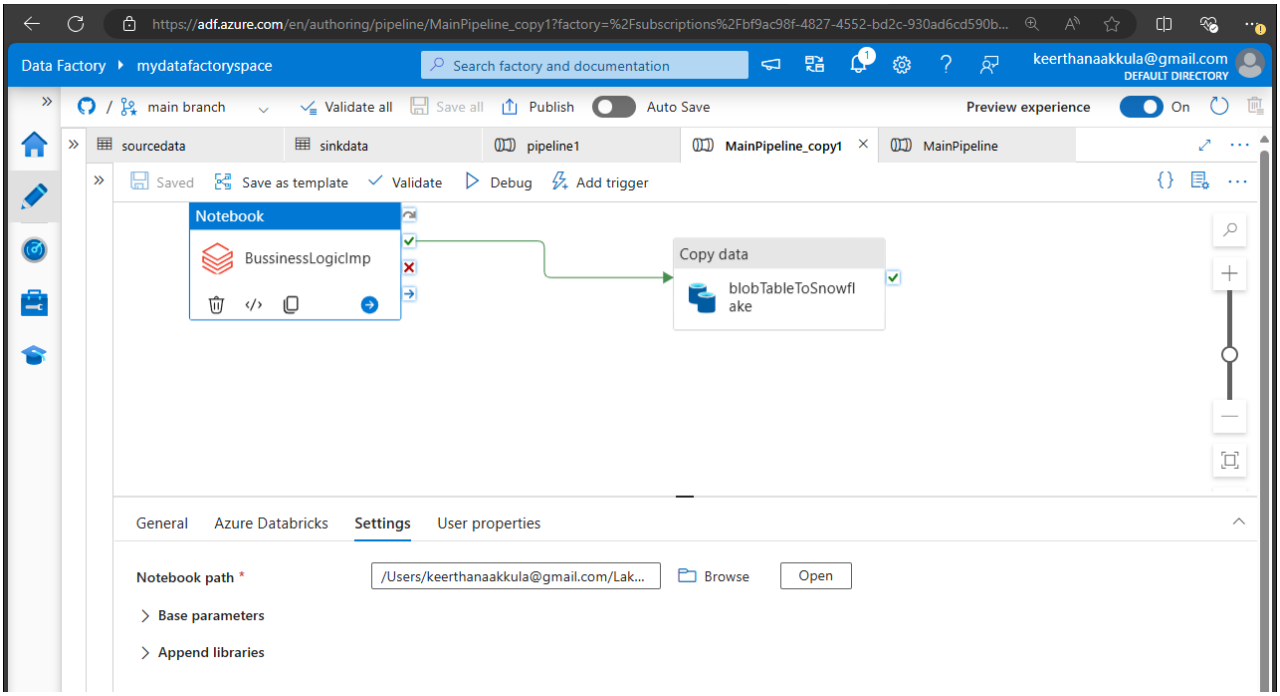

InPrivate (2)
mydatafactoryspace - Azure Dat
container - Microsoft Azure
https://adf.azure.com/en/authoring/dataset/SnowflakeTable1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2F...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Snowflake
SnowflakeTable1
Connection
Schema
Parameters
Import schema
Clear
Column name
Type
CUSTOMERID
NUMBER
YEAR
NUMBER
MONTH
NUMBER
TOTALPURCHASES
NUMBER
AVERAGEPURCHASEQUANTITY
FLOAT

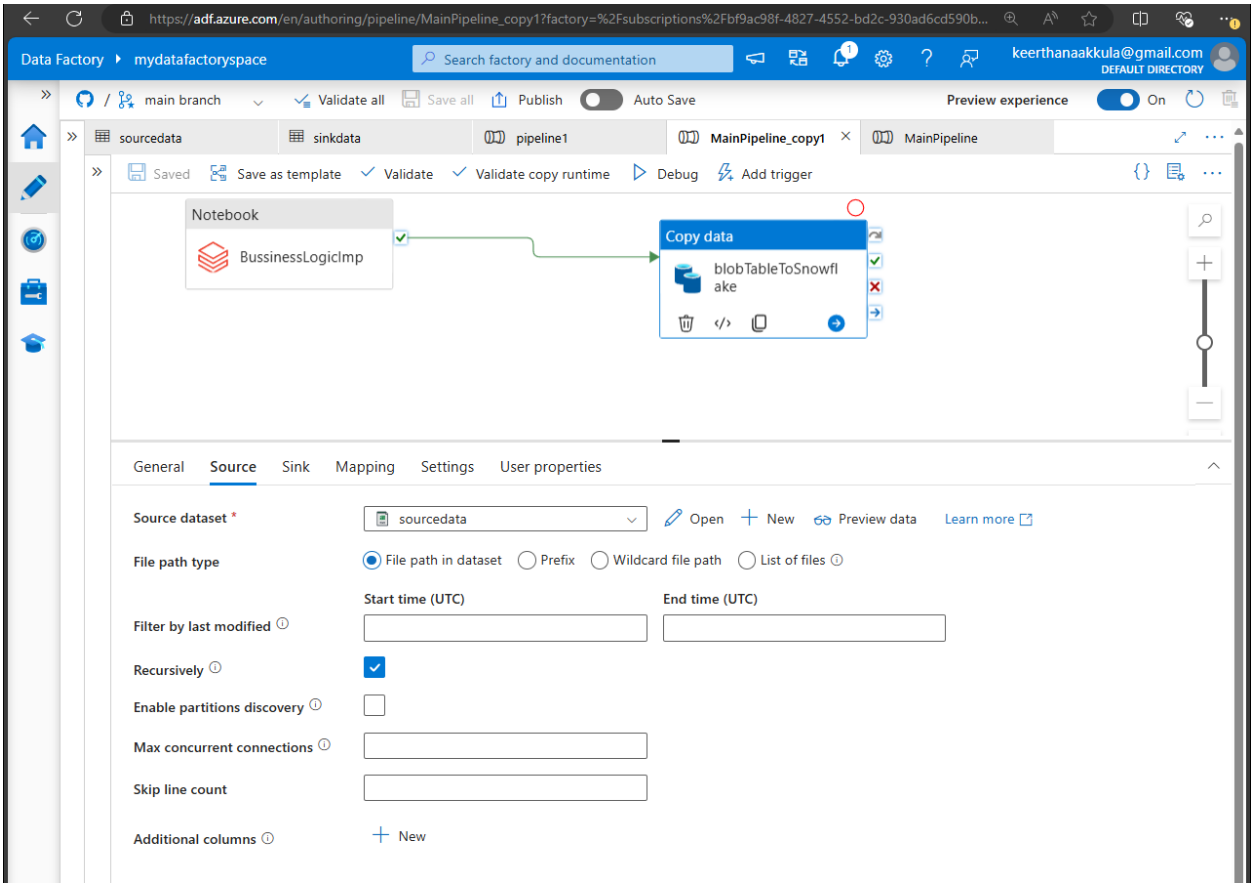

https://adf.azure.com/en/authoring/pipeline/MainPipeline_copy1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b...
Data Factory
Validate all
Publish all 2
Auto Save
Preview experience
On
AzureDatabricksDelt...
SnowflakeTable1
MainPipeline
MainPipeline_copy1
Activities
Search activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Notebook
BussinessLogicImp
Copy data
deltaTableToSnowflake
General
Source
Sink
Mapping
Settings
User properties
Import schemas
Preview source
New mapping
Clear
Reset
Delete
Source
Type
Destination
Type
CustomerID
integer
CUSTOMERID
NUMBER
Year
integer
YEAR
NUMBER
Month
integer
MONTH
NUMBER
TotalPurchases
long
TOTALPURCHASES
NUMBER
AveragePurchaseQua...
double
AVERAGEPURCHASEQ...
FLOAT
Add dynamic content [Alt+Shift+D]

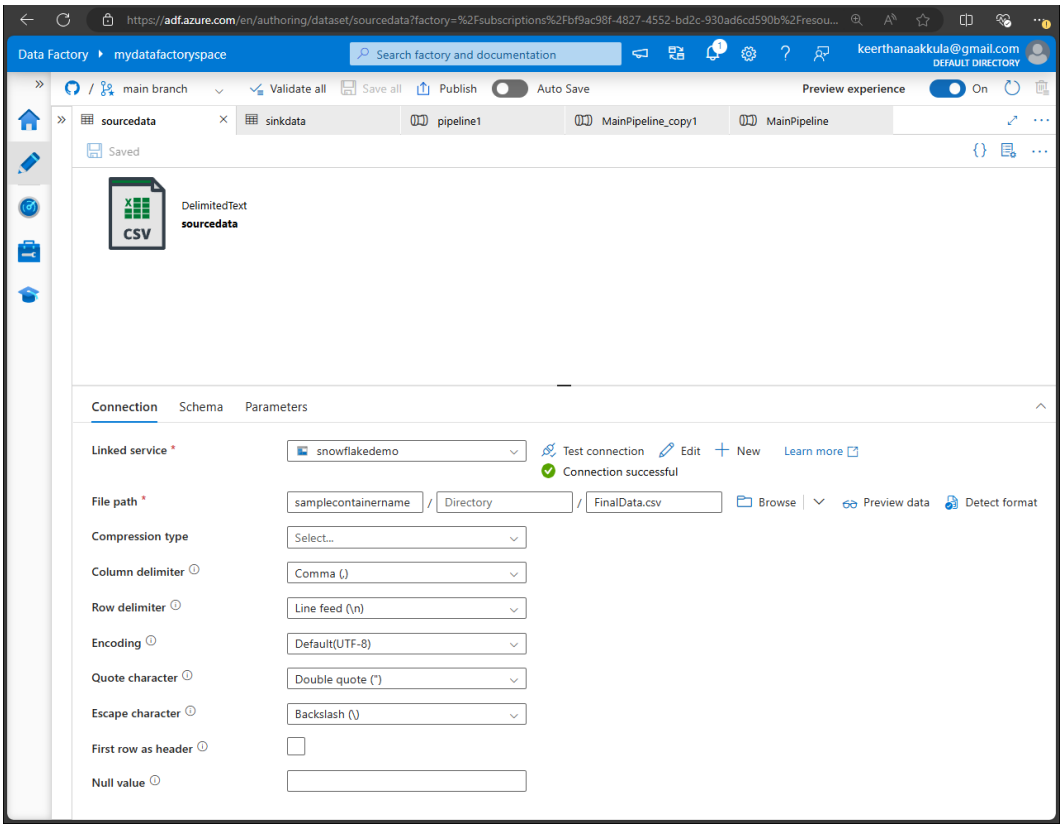
The above tried approach is loading the data directly from delta lake table (DATABRICKS) to snowflake in data factory using linked services (SAS URL)

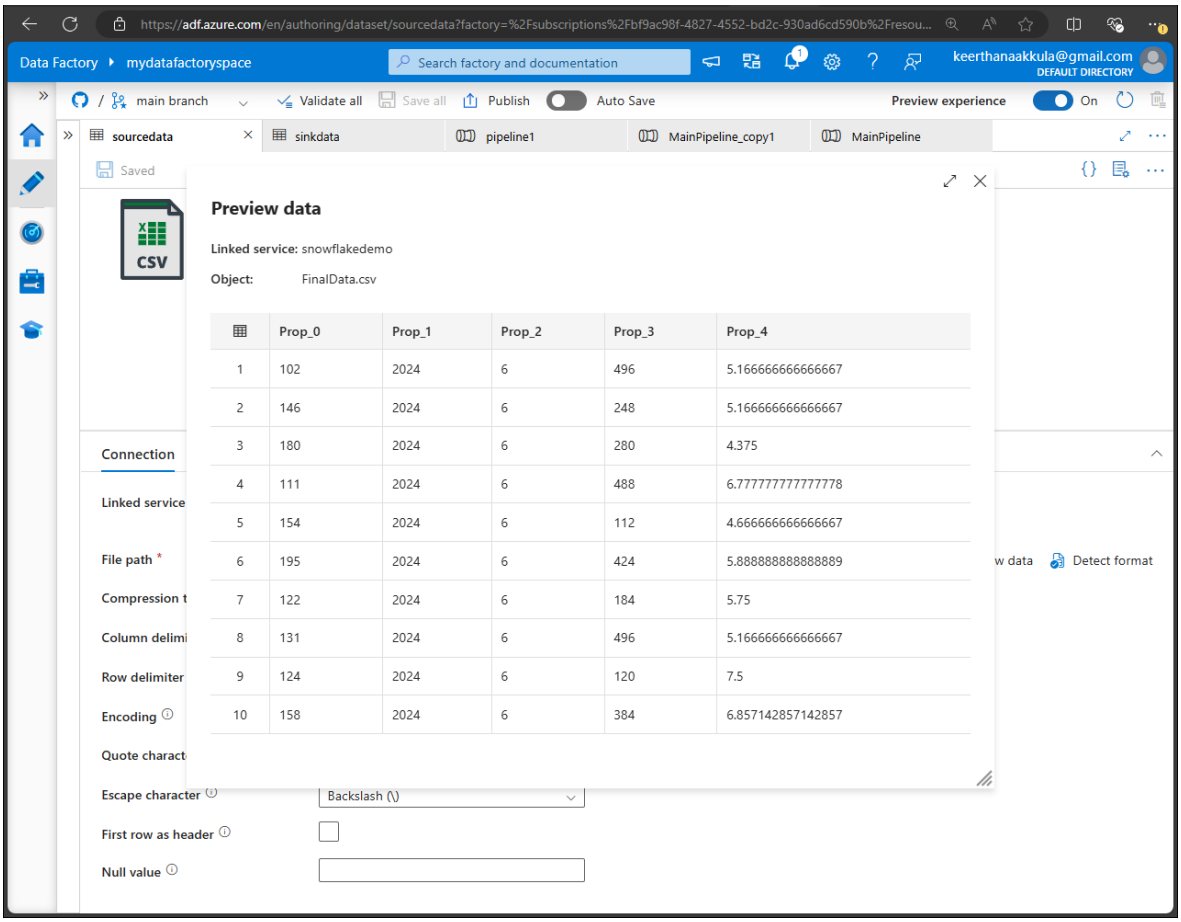
Alternative approach to load the data from blob storage(csv) to snowflake

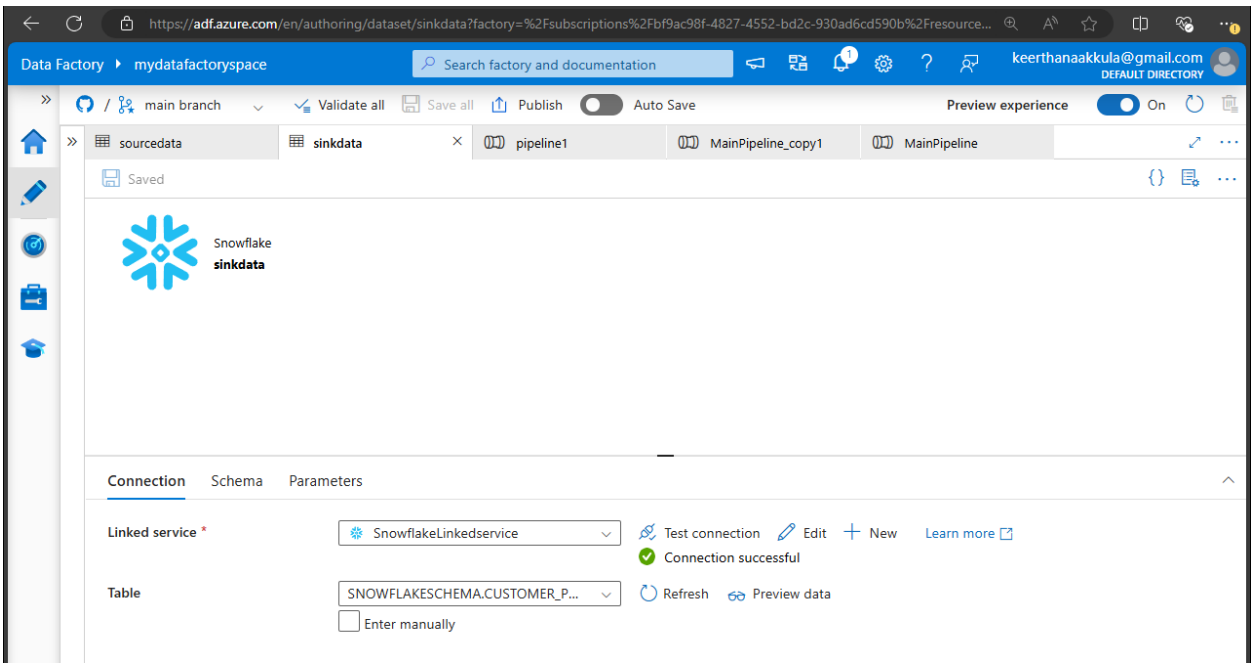


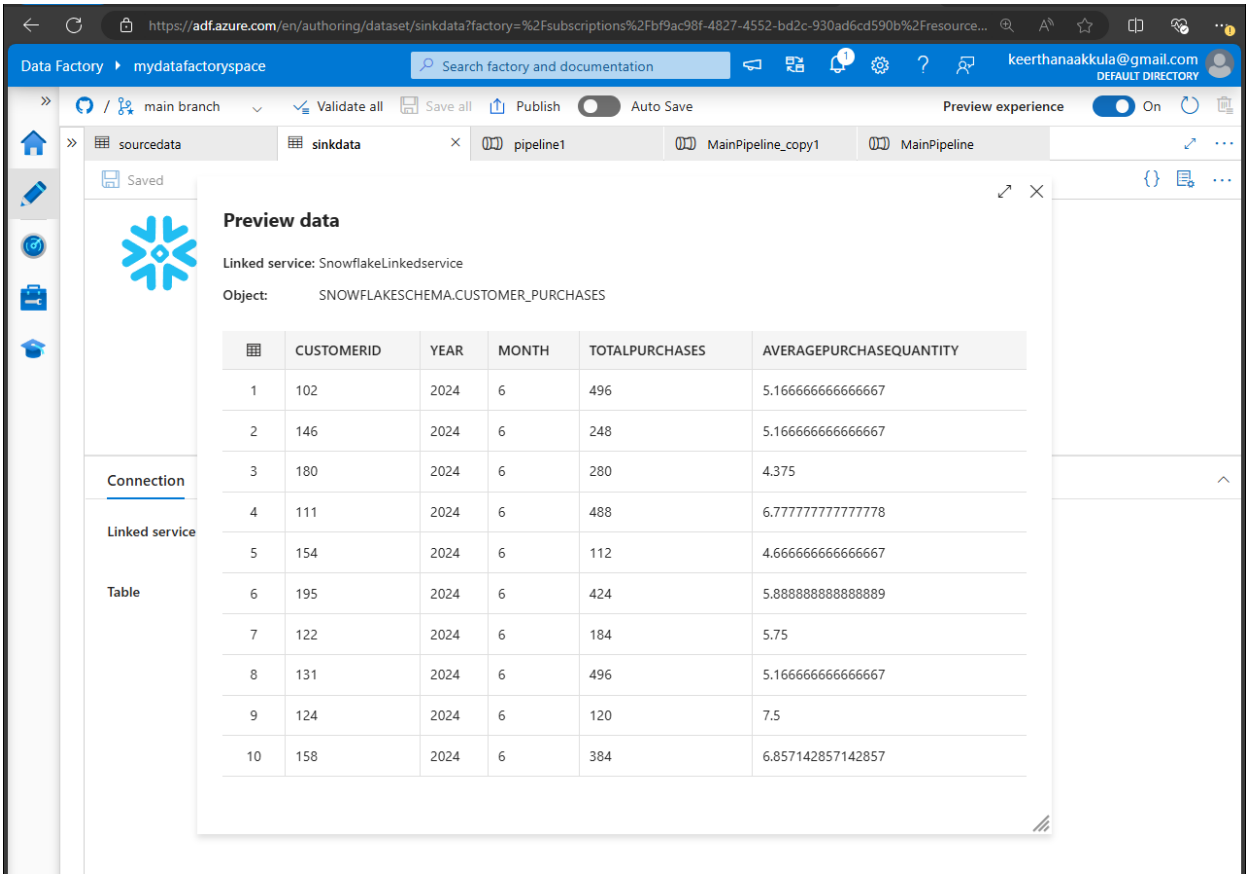




https://adf.azure.com/en/authoring/pipeline/MainPipeline_copy1?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b...
Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
main branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
sourcedata
sinkdata
pipeline1
MainPipeline_copy1
MainPipeline
Saved
Save as template
Validate
Validate copy runtime
Debug
Add trigger
Notebook
BussinessLogicImp
Copy data
blobTableToSnowflake
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
sourcedata
Open
New
Preview data
Learn more
File path type
File path in dataset
Prefix
Wildcard file path
List of files
Start time (UTC)
End time (UTC)
Filter by last modified
Recursively
Enable partitions discovery
Max concurrent connections
Skip line count
Additional columns
New

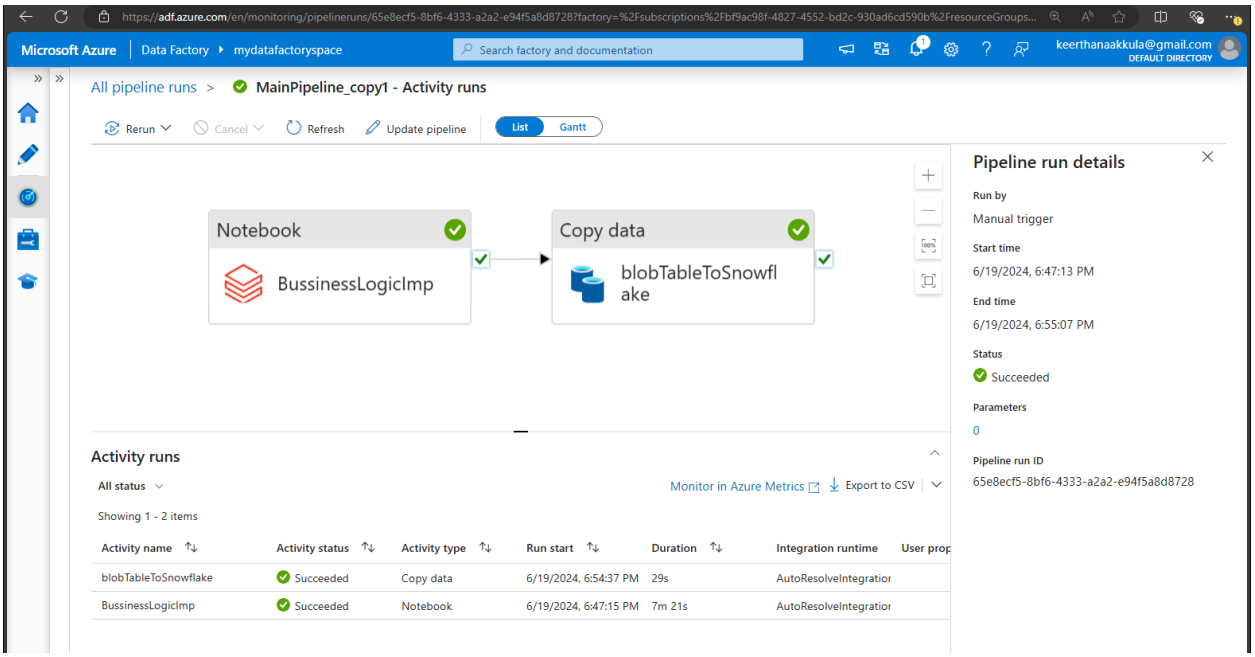

https://adf.azure.com/en/authoring/dataset/sourcedata?factory=%2Fsubscriptions%2Fbf9ac98f-4827-4552-bd2c-930ad6cd590b%2Fresou...
Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
main branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
sourcedata
sinkdata
pipeline1
MainPipeline_copy1
MainPipeline
Saved
CSV
DelimitedText
sourcedata
Connection
Schema
Parameters
Linked service *
snowflakedemo
Test connection
Edit
New
Learn more
Connection successful
File path *
samplecontainername
Directory
FinalData.csv
Browse
Preview data
Detect format
Compression type
Select...
Column delimiter
Comma (,)
Row delimiter
Line feed (\n)
Encoding
Default(UTF-8)
Quote character
Double quote (")
Escape character
Backslash (\)
First row as header
Null value

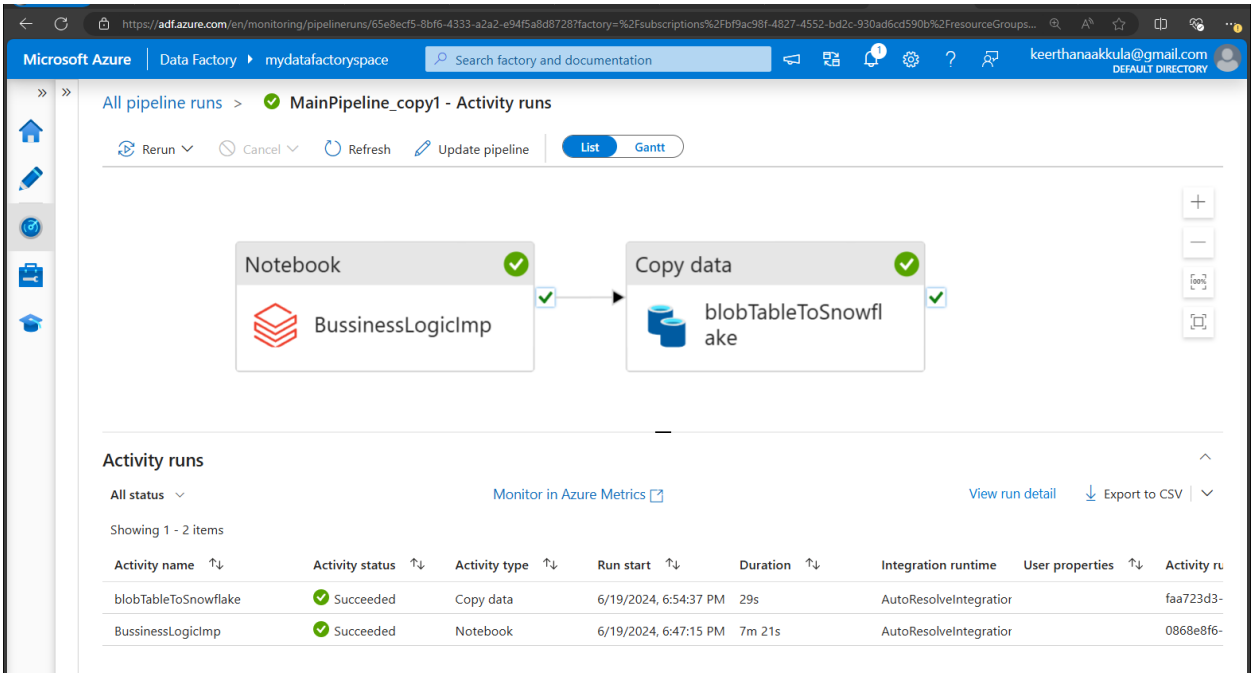

Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
main branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
sourcedata
sinkdata
pipeline1
MainPipeline_copy1
MainPipeline
Saved
CSV
Preview data
Linked service: snowflakedemo
Object: FinalData.csv
Prop_0 Prop_1 Prop_2 Prop_3 Prop_4
1 102 2024 6 496 5.166666666666667
2 146 2024 6 248 5.166666666666667
3 180 2024 6 280 4.375
4 111 2024 6 488 6.777777777777778
5 154 2024 6 112 4.666666666666667
6 195 2024 6 424 5.888888888888889
7 122 2024 6 184 5.75
8 131 2024 6 496 5.166666666666667
9 124 2024 6 120 7.5
10 158 2024 6 384 6.857142857142857
Connection
Linked service
File path *
Compression t
Column delimi
Row delimiter
Encoding
Quote charact
Escape character
Backslash (\)
First row as header
Null value
w data
Detect format

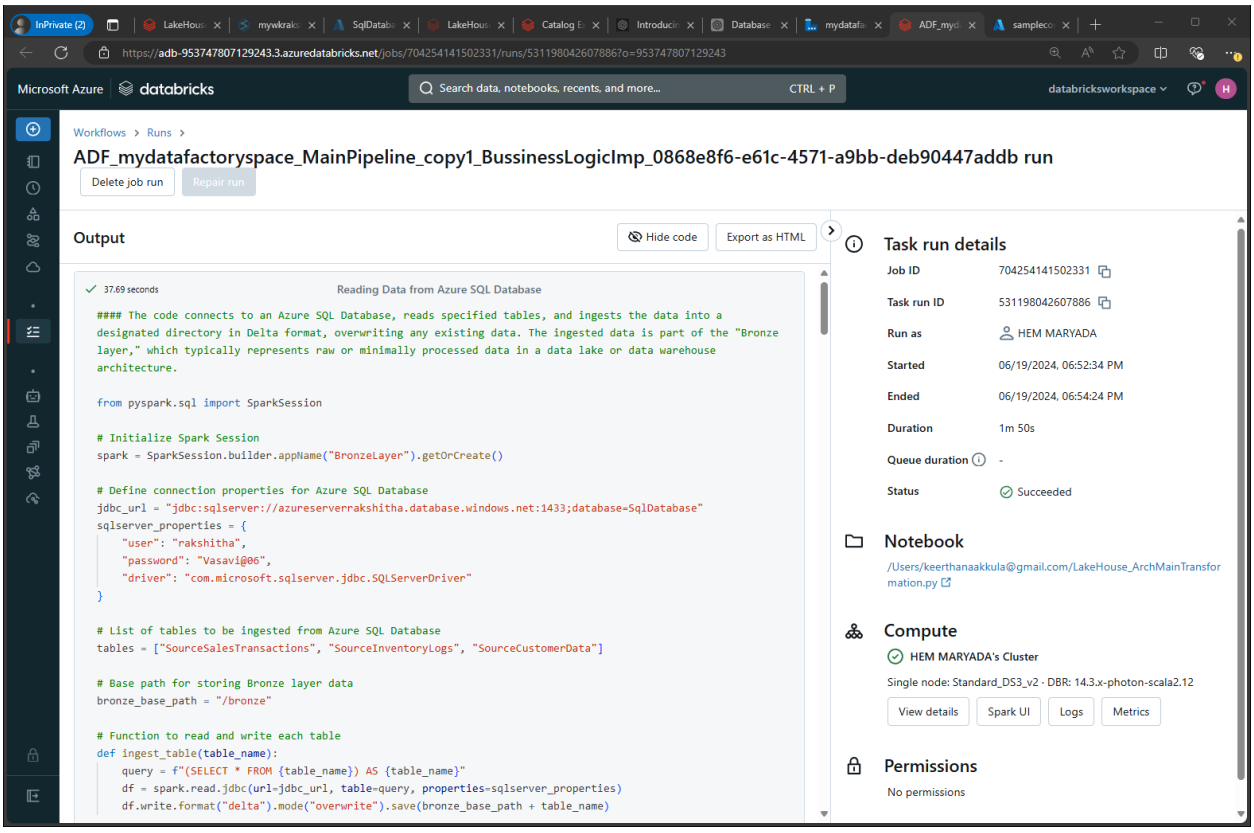

Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
main branch
Validate all
Save all
Publish
Auto Save
Preview experience
On
sourcedata
sinkdata
pipeline1
MainPipeline_copy1
MainPipeline
Saved
Snowflake
sinkdata
Connection
Schema
Parameters
Linked service *
SnowflakeLinkedservice
Test connection
Edit
New
Learn more
Connection successful
Table
SNOWFLAKESCHEMA.CUSTOMER_P...
Refresh
Preview data
Enter manually



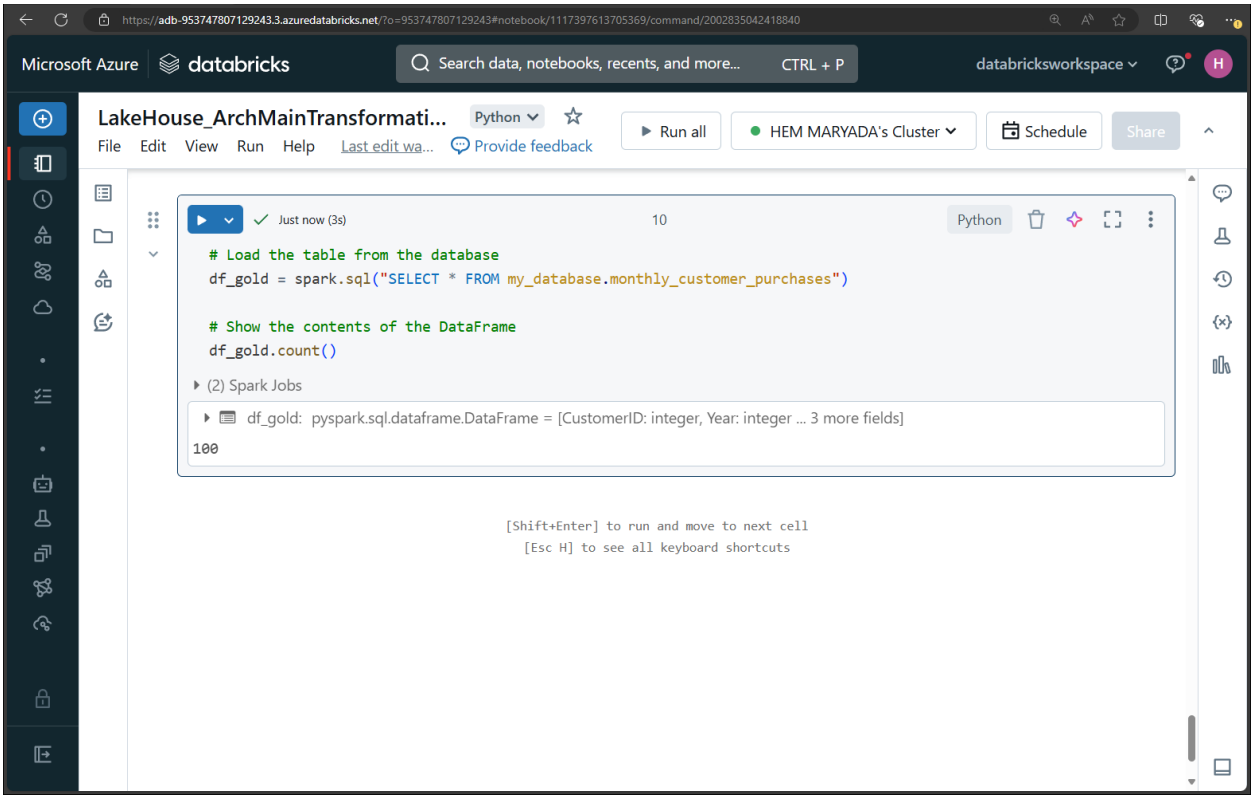

| | CUSTOMERID | YEAR | MONTH | TOTALPURCHASES | AVERAGEPURCHASEQUANTITY |
|---|---|---|---|---|---|
| 1 | 102 | 2024 | 6 | 496 | 5.166666666666667 |
| 2 | 146 | 2024 | 6 | 248 | 5.166666666666667 |
| 3 | 180 | 2024 | 6 | 280 | 4.375 |
| 4 | 111 | 2024 | 6 | 488 | 6.777777777777778 |
| 5 | 154 | 2024 | 6 | 112 | 4.666666666666667 |
| 6 | 195 | 2024 | 6 | 424 | 5.888888888888889 |
| 7 | 122 | 2024 | 6 | 184 | 5.75 |
| 8 | 131 | 2024 | 6 | 496 | 5.166666666666667 |
| 9 | 124 | 2024 | 6 | 120 | 7.5 |
| 10 | 158 | 2024 | 6 | 384 | 6.857142857142857 |

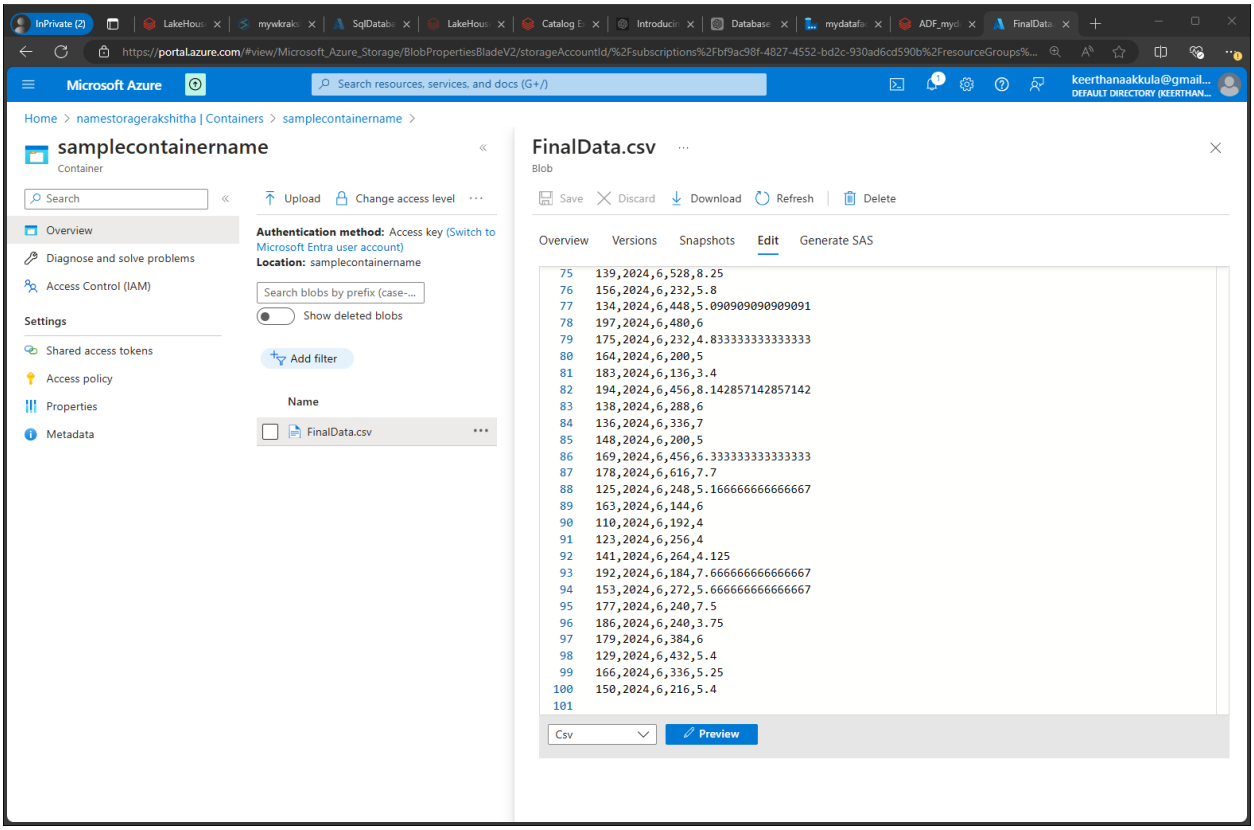


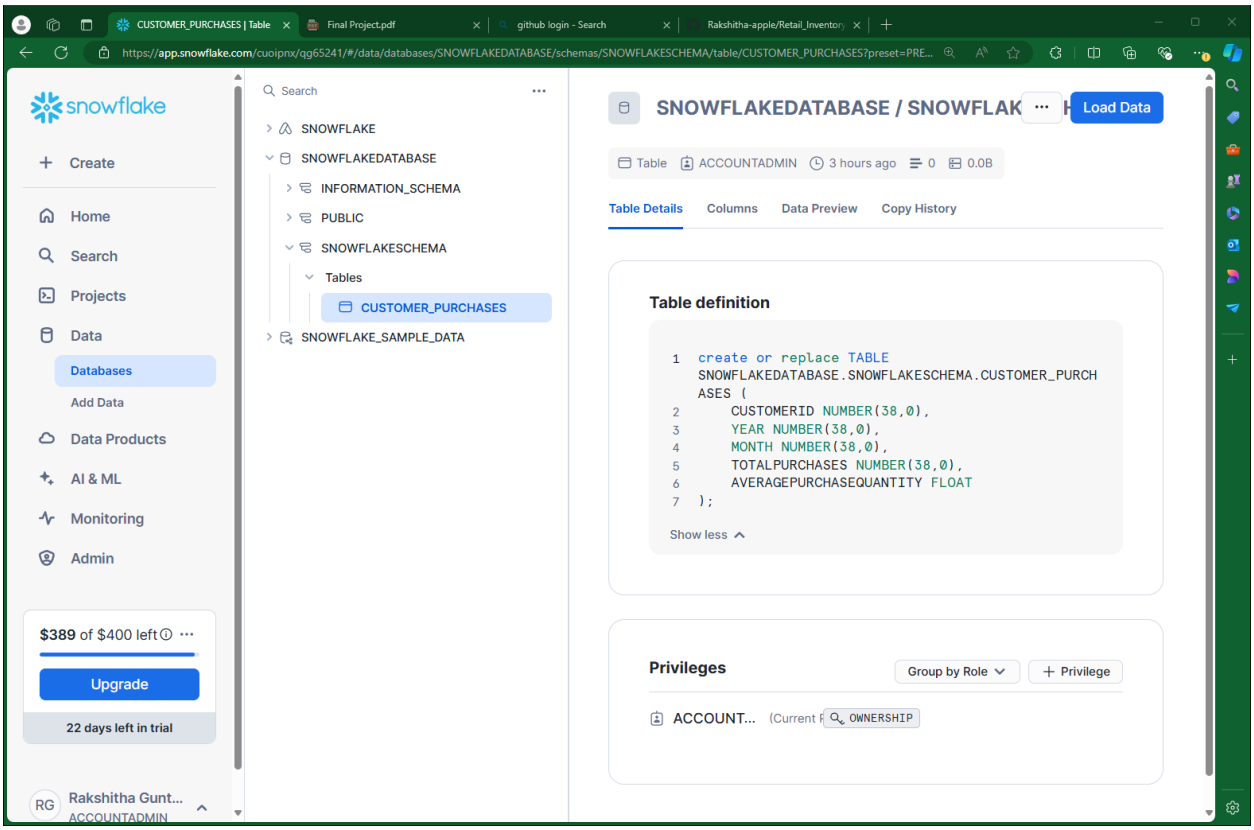

Microsoft Azure
Data Factory
mydatafactoryspace
Search factory and documentation
keerthanaakkula@gmail.com
DEFAULT DIRECTORY
All pipeline runs
MainPipeline_copy1 - Activity runs
Rerun
Cancel
Refresh
Update pipeline
List
Gantt
Notebook
BussinessLogicImp
Copy data
blobTableToSnowflake
Activity runs
All status
Monitor in Azure Metrics
View run detail
Export to CSV
Showing 1 - 2 items
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
User properties
blobTableToSnowflake
Succeeded
Copy data
6/19/2024, 6:54:37 PM
29s
AutoResolveIntegratior
faa723d3-
BussinessLogicImp
Succeeded
Notebook
6/19/2024, 6:47:15 PM
7m 21s
AutoResolveIntegratior
0868e8f6-

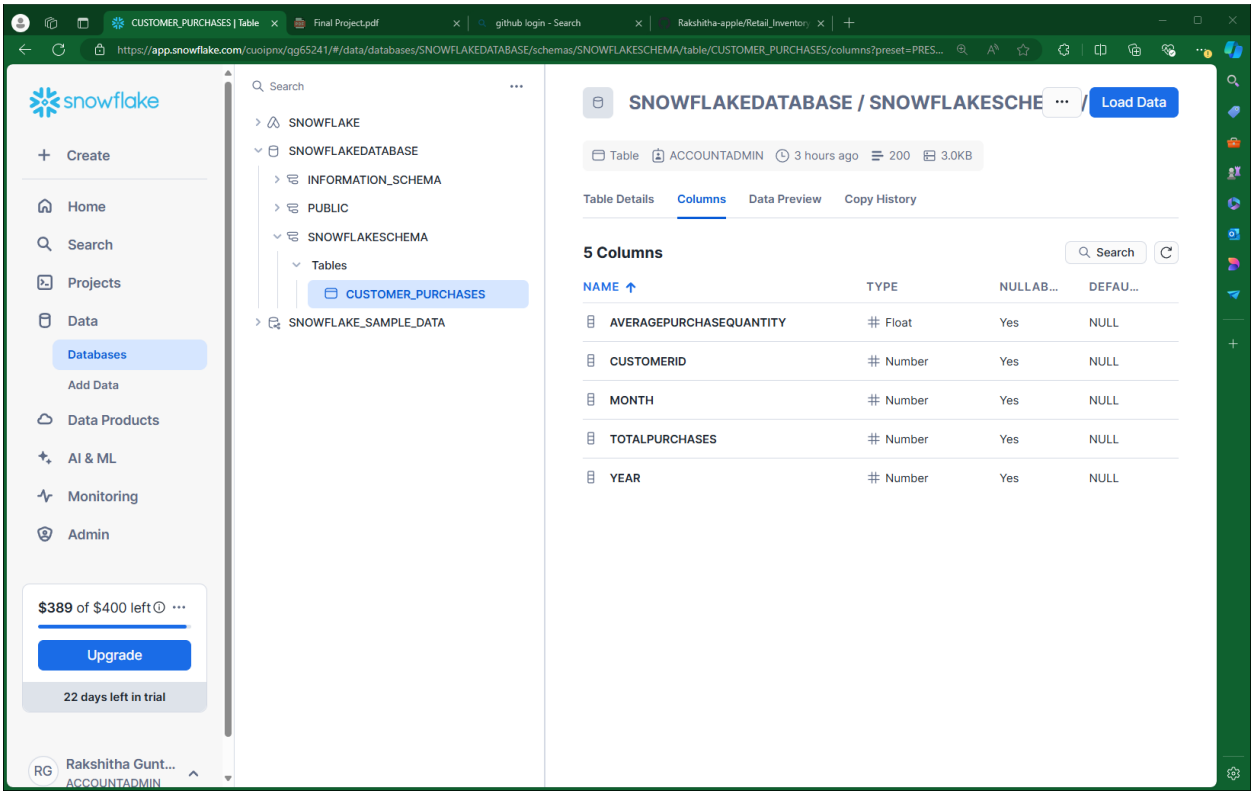

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
databricksworkspace
Workflows
Runs
ADF_mydatafactoryspace_MainPipeline_copy1_BussinessLogicImp_0868e8f6-e61c-4571-a9bb-deb90447addb run
Delete job run
Repair run
Output
Hide code
Export as HTML
37.69 seconds
Reading Data from Azure SQL Database
#### The code connects to an Azure SQL Database, reads specified tables, and ingests the data into a designated directory in Delta format, overwriting any existing data. The ingested data is part of the "Bronze layer," which typically represents raw or minimally processed data in a data lake or data warehouse architecture.
from pyspark.sql import SparkSession
# Initialize Spark Session
spark = SparkSession.builder.appName("BronzeLayer").getOrCreate()
# Define connection properties for Azure SQL Database
jdbc_url = "jdbc:sqlserver://azureserverrakshitha.database.windows.net:1433;database=SqlDatabase"
sqlserver_properties = {
"user": "rakshitha",
"password": "Vasavi@06",
"driver": "com.microsoft.sqlserver.jdbc.SQLServerDriver"
}
# List of tables to be ingested from Azure SQL Database
tables = ["SourceSalesTransactions", "SourceInventoryLogs", "SourceCustomerData"]
# Base path for storing Bronze layer data
bronze_base_path = "/bronze"
# Function to read and write each table
def ingest_table(table_name):
query = f"(SELECT * FROM {table_name}) AS {table_name}"
df = spark.read.jdbc(url=jdbc_url, table=query, properties=sqlserver_properties)
df.write.format("delta").mode("overwrite").save(bronze_base_path + table_name)
Task run details
Job ID
704254141502331
Task run ID
531198042607886
Run as
HEM MARYADA
Started
06/19/2024, 06:52:34 PM
Ended
06/19/2024, 06:54:24 PM
Duration
1m 50s
Queue duration
-
Status
Succeeded
Notebook
/Users/keerthanaakkula@gmail.com/LakeHouse_ArchMainTransformation.py
Compute
HEM MARYADA's Cluster
Single node: Standard_DS3_v2 · DBR: 14.3.x-photon-scala2.12
View details
Spark UI
Logs
Metrics
Permissions
No permissions

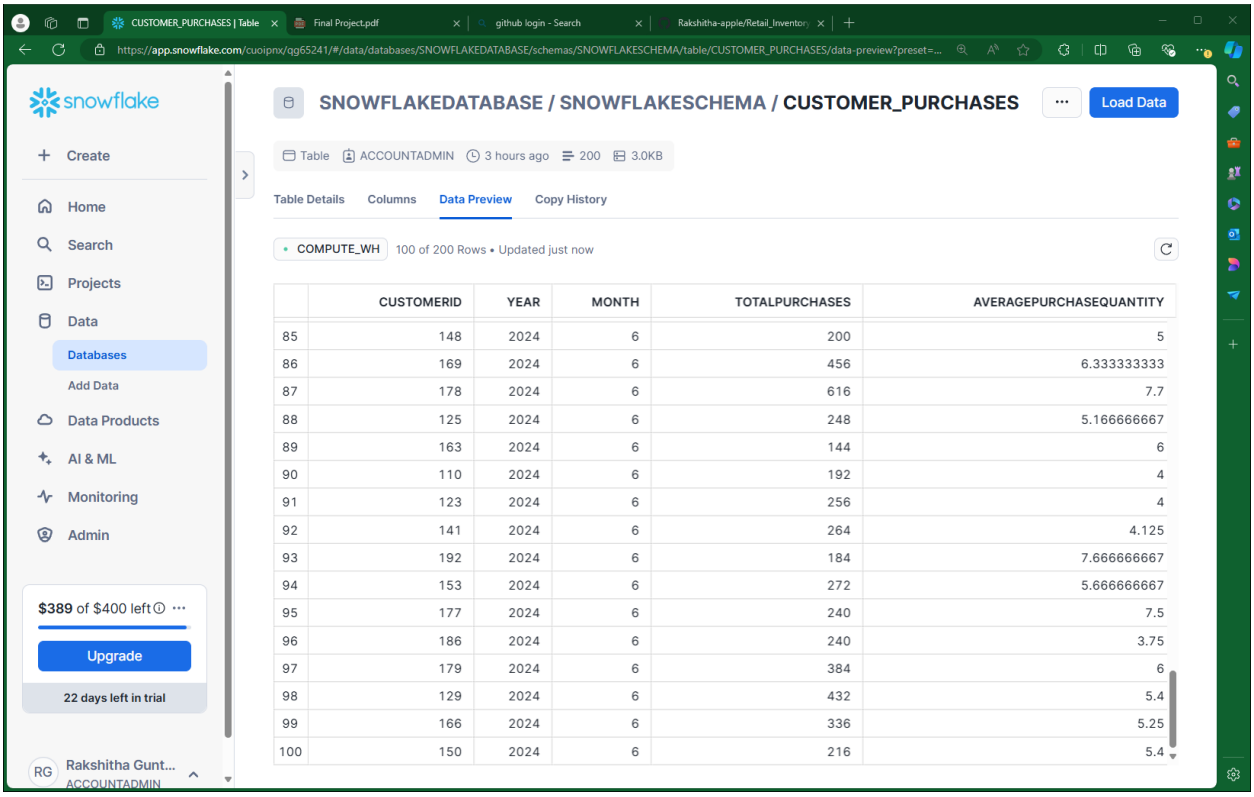
Count: 100


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
databricksworkspace
LakeHouse_ArchMainTransformati...
Python
File Edit View Run Help
Provide feedback
Run all
HEM MARYADA's Cluster
Schedule
Share
Just now (3s)
10
Python
# Load the table from the database
df_gold = spark.sql("SELECT * FROM my_database.monthly_customer_purchases")
# Show the contents of the DataFrame
df_gold.count()
(2) Spark Jobs
df_gold: pyspark.sql.dataframe.DataFrame = [CustomerID: integer, Year: integer ... 3 more fields]
100
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

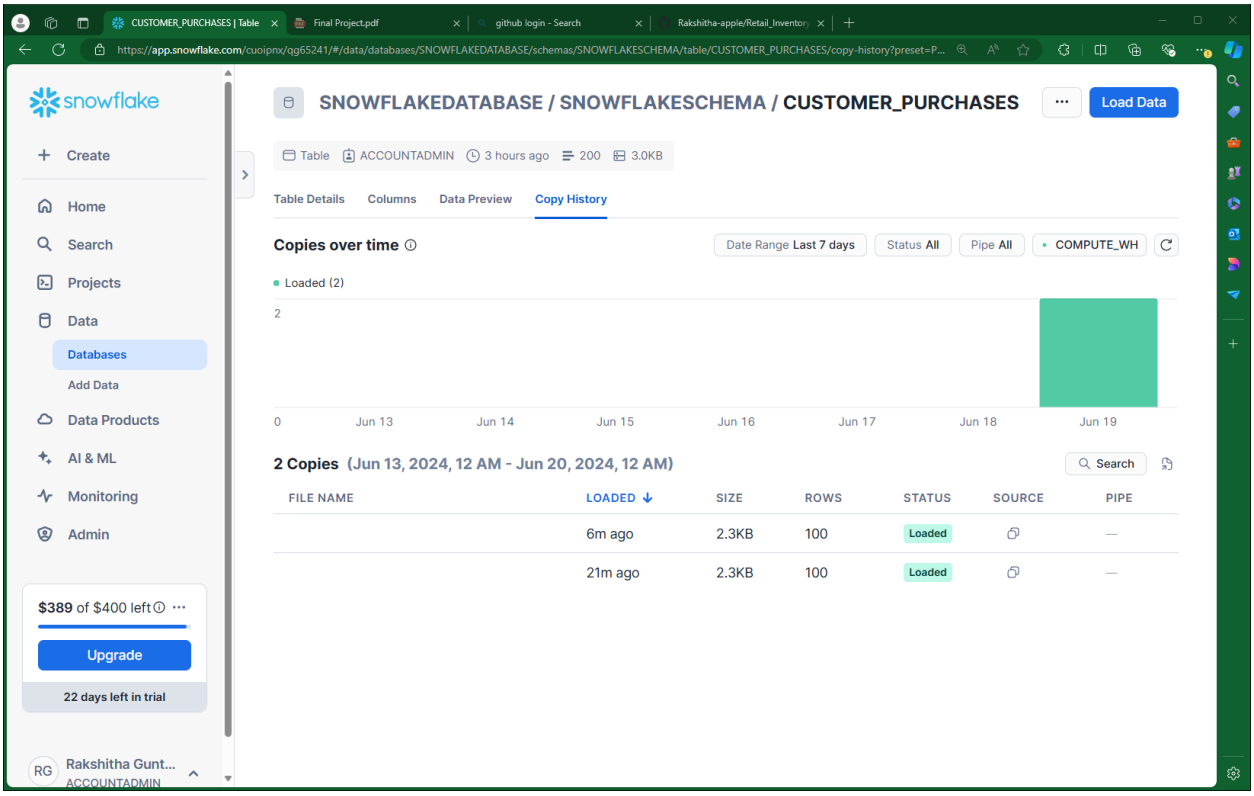

Microsoft Azure
Search resources, services, and docs (G+/)
Home > namestoragerakshitha | Containers > samplecontainername >
samplecontainername
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Access policy
Properties
Metadata
Upload
Change access level
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: samplecontainername
Search blobs by prefix (case-...
Show deleted blobs
Add filter
Name
FinalData.csv
FinalData.csv
Blob
Save
Discard
Download
Refresh
Delete
Overview
Versions
Snapshots
Edit
Generate SAS
75 139,2024,6,528,8.25
76 156,2024,6,232,5.8
77 134,2024,6,448,5.090909090909091
78 197,2024,6,480,6
79 175,2024,6,232,4.833333333333333
80 164,2024,6,200,5
81 183,2024,6,136,3.4
82 194,2024,6,456,8.142857142857142
83 138,2024,6,288,6
84 136,2024,6,336,7
85 148,2024,6,200,5
86 169,2024,6,456,6.333333333333333
87 178,2024,6,616,7.7
88 125,2024,6,248,5.166666666666667
89 163,2024,6,144,6
90 110,2024,6,192,4
91 123,2024,6,256,4
92 141,2024,6,264,4.125
93 192,2024,6,184,7.666666666666667
94 153,2024,6,272,5.666666666666667
95 177,2024,6,240,7.5
96 186,2024,6,240,3.75
97 179,2024,6,384,6
98 129,2024,6,432,5.4
99 166,2024,6,336,5.25
100 150,2024,6,216,5.4
101
Csv
Preview


CUSTOMER_PURCHASES | Table
Final Project.pdf
github login - Search
Rakshitha-apple/Retail_Inventory
https://app.snowflake.com/cuoipnx/qg65241/#/data/databases/SNOWFLAKEDATABASE/schemas/SNOWFLAKESCHEMA/table/CUSTOMER_PURCHASES?preset=PRE...
snowflake
Create
Home
Search
Projects
Data
Databases
Add Data
Data Products
AI & ML
Monitoring
Admin
$389 of $400 left
Upgrade
22 days left in trial
Rakshitha Gunt...
ACCOUNTADMIN
Search
SNOWFLAKE
SNOWFLAKEDATABASE
INFORMATION_SCHEMA
PUBLIC
SNOWFLAKESCHEMA
Tables
CUSTOMER_PURCHASES
SNOWFLAKE_SAMPLE_DATA
SNOWFLAKEDATABASE / SNOWFLAK
Load Data
Table
ACCOUNTADMIN
3 hours ago
0
0.0B
Table Details
Columns
Data Preview
Copy History
Table definition
1 create or replace TABLE SNOWFLAKEDATABASE.SNOWFLAKESCHEMA.CUSTOMER_PURCHASES (
2 CUSTOMERID NUMBER(38,0),
3 YEAR NUMBER(38,0),
4 MONTH NUMBER(38,0),
5 TOTALPURCHASES NUMBER(38,0),
6 AVERAGEPURCHASEQUANTITY FLOAT
7 );
Show less
Privileges
Group by Role
Privilege
ACCOUNT...
(Current
OWNERSHIP


CUSTOMER_PURCHASES | Table
Final Project.pdf
github login - Search
Rakshitha-apple/Retail_Inventory
https://app.snowflake.com/cuoipnx/qg65241/#/data/databases/SNOWFLAKEDATABASE/schemas/SNOWFLAKESCHEMA/table/CUSTOMER_PURCHASES/columns?preset=PRES...
snowflake
Create
Home
Search
Projects
Data
Databases
Add Data
Data Products
AI & ML
Monitoring
Admin
$389 of $400 left
Upgrade
22 days left in trial
Rakshitha Gunt...
ACCOUNTADMIN
Search
SNOWFLAKE
SNOWFLAKEDATABASE
INFORMATION_SCHEMA
PUBLIC
SNOWFLAKESCHEMA
Tables
CUSTOMER_PURCHASES
SNOWFLAKE_SAMPLE_DATA
SNOWFLAKEDATABASE / SNOWFLAKESCHE
Load Data
Table
ACCOUNTADMIN
3 hours ago
200
3.0KB
Table Details
Columns
Data Preview
Copy History
5 Columns
Search
NAME
TYPE
NULLAB...
DEFAU...
AVERAGEPURCHASEQUANTITY
Float
Yes
NULL
CUSTOMERID
Number
Yes
NULL
MONTH
Number
Yes
NULL
TOTALPURCHASES
Number
Yes
NULL
YEAR
Number
Yes
NULL




| | CUSTOMERID | YEAR | MONTH | TOTALPURCHASES | AVERAGEPURCHASEQUANTITY |
|---|---|---|---|---|---|
| 85 | 148 | 2024 | 6 | 200 | 5 |
| 86 | 169 | 2024 | 6 | 456 | 6.333333333 |
| 87 | 178 | 2024 | 6 | 616 | 7.7 |
| 88 | 125 | 2024 | 6 | 248 | 5.166666667 |
| 89 | 163 | 2024 | 6 | 144 | 6 |
| 90 | 110 | 2024 | 6 | 192 | 4 |
| 91 | 123 | 2024 | 6 | 256 | 4 |
| 92 | 141 | 2024 | 6 | 264 | 4.125 |
| 93 | 192 | 2024 | 6 | 184 | 7.666666667 |
| 94 | 153 | 2024 | 6 | 272 | 5.666666667 |
| 95 | 177 | 2024 | 6 | 240 | 7.5 |
| 96 | 186 | 2024 | 6 | 240 | 3.75 |
| 97 | 179 | 2024 | 6 | 384 | 6 |
| 98 | 129 | 2024 | 6 | 432 | 5.4 |
| 99 | 166 | 2024 | 6 | 336 | 5.25 |
| 100 | 150 | 2024 | 6 | 216 | 5.4 |

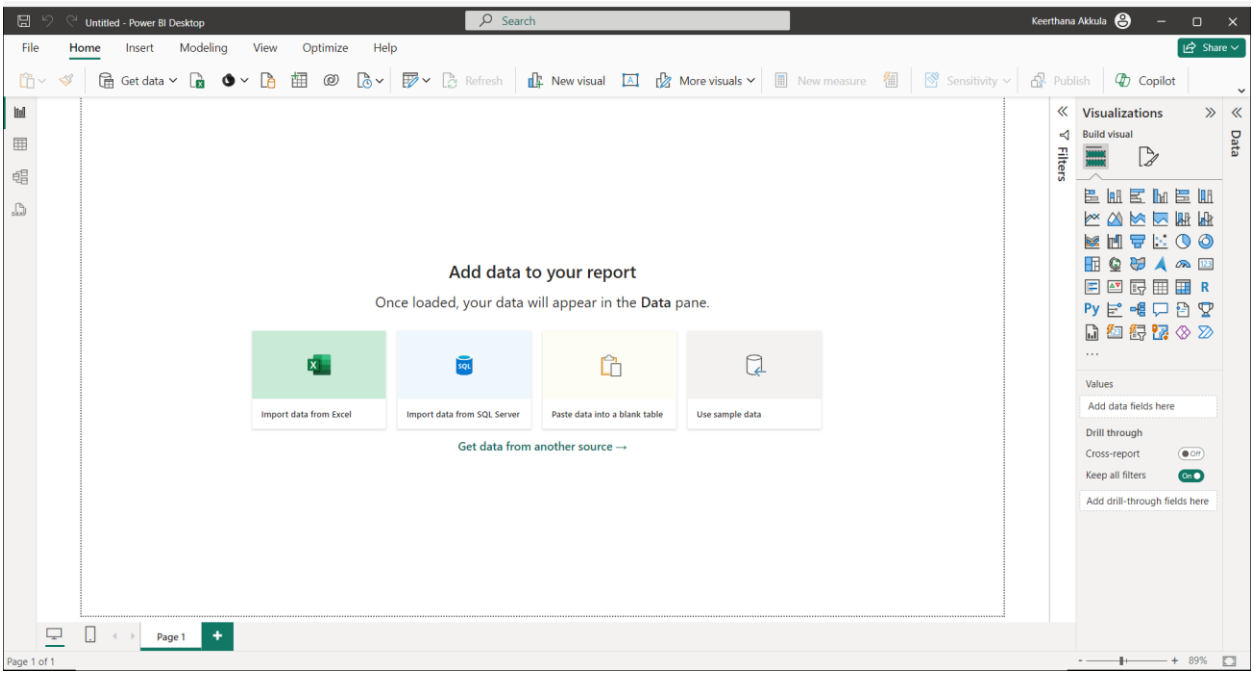


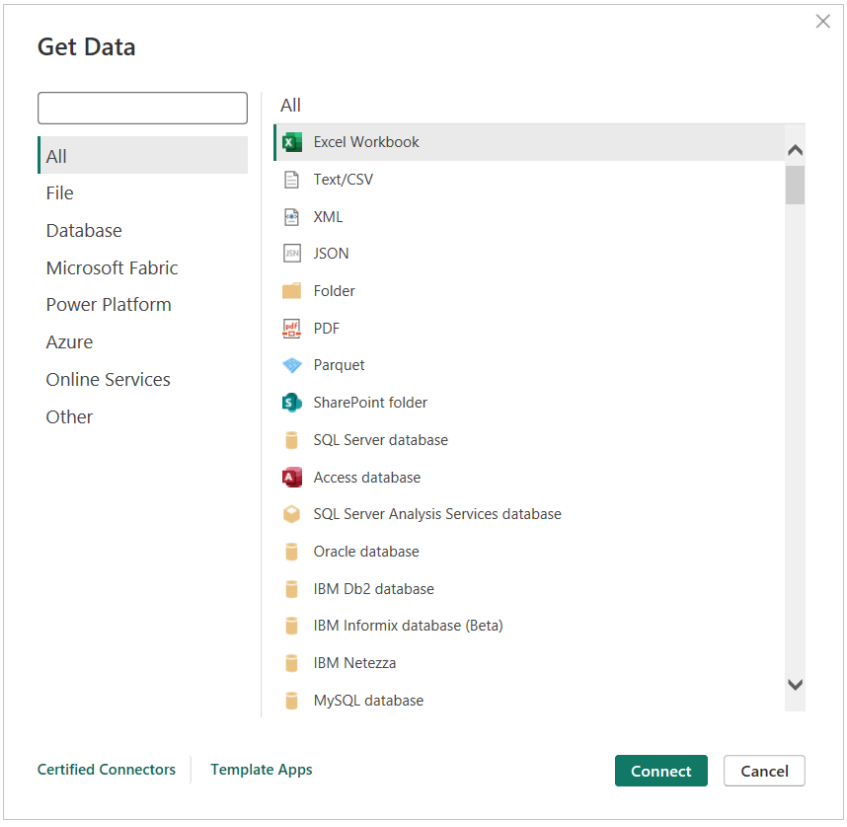

| FILE NAME | LOADED ↓ | SIZE | ROWS | STATUS | SOURCE | PIPE |
|---|---|---|---|---|---|---|
| | 6m ago | 2.3KB | 100 | Loaded | | — |
| | 21m ago | 2.3KB | 100 | Loaded | | — |

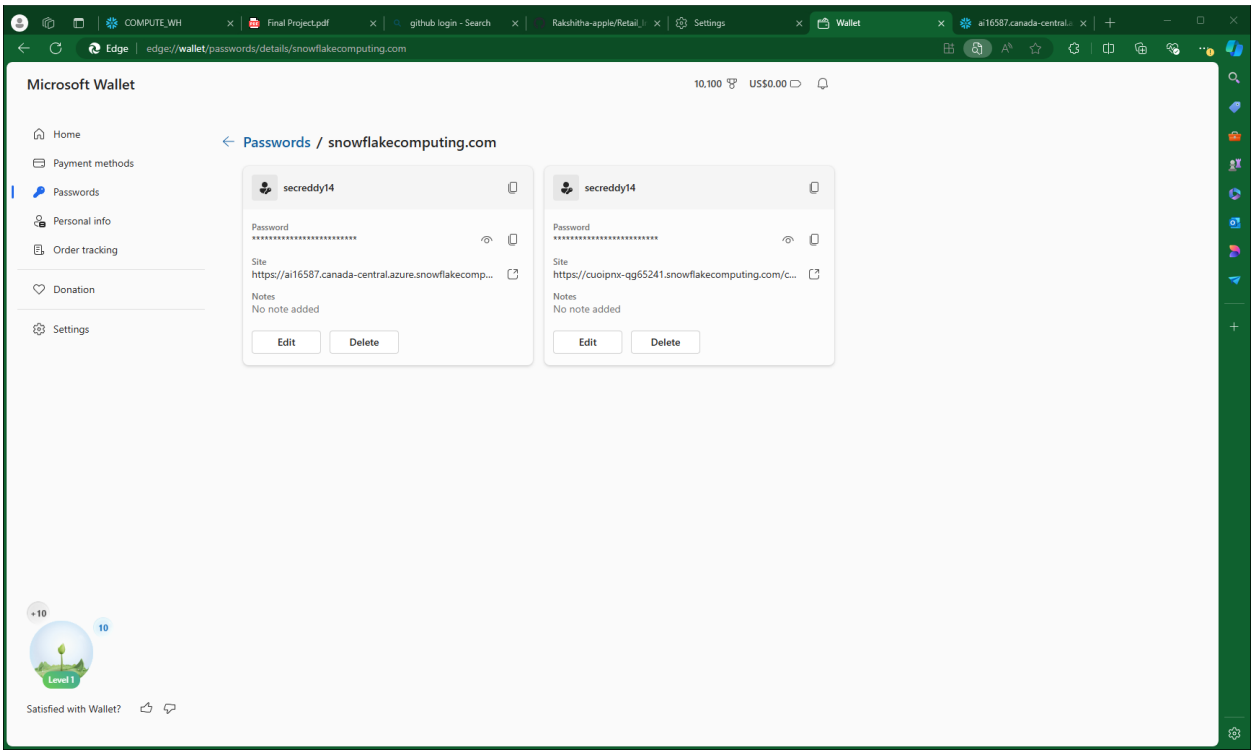
**Snowflake to Power BI Desktop**

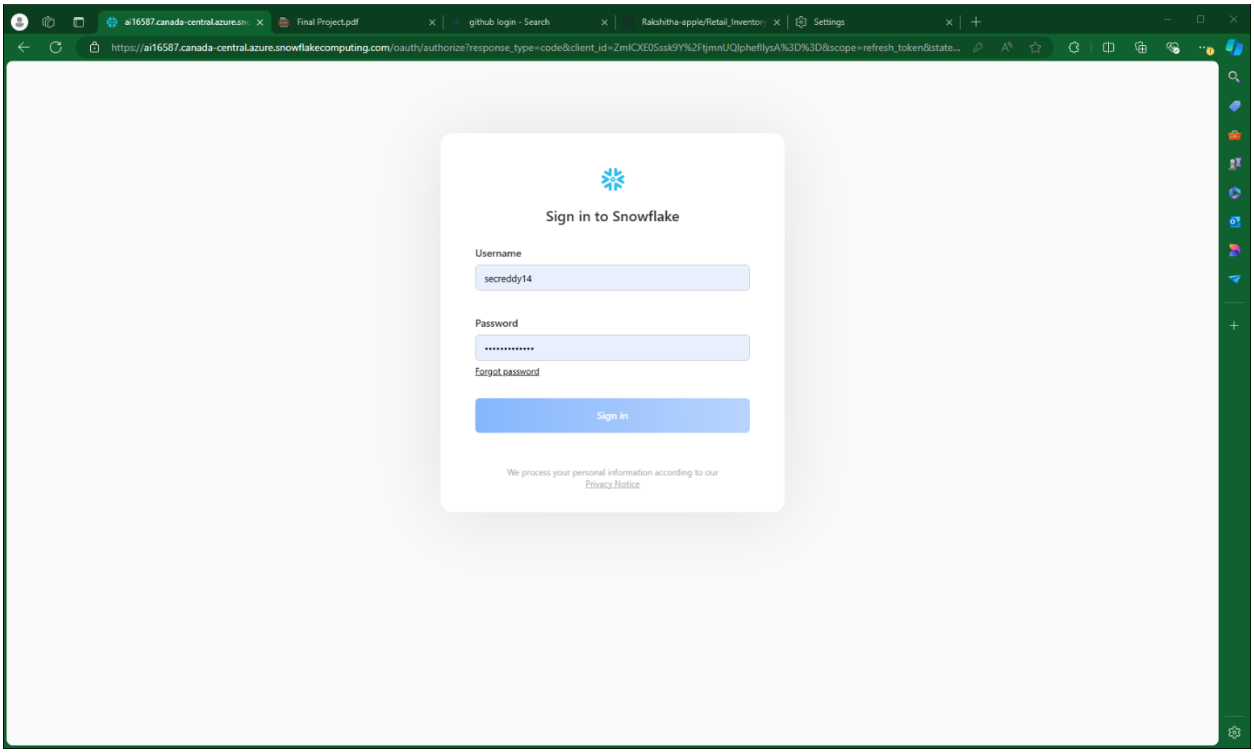
**Server name of snowflake Accountname.region.snowflakecomputing.com**

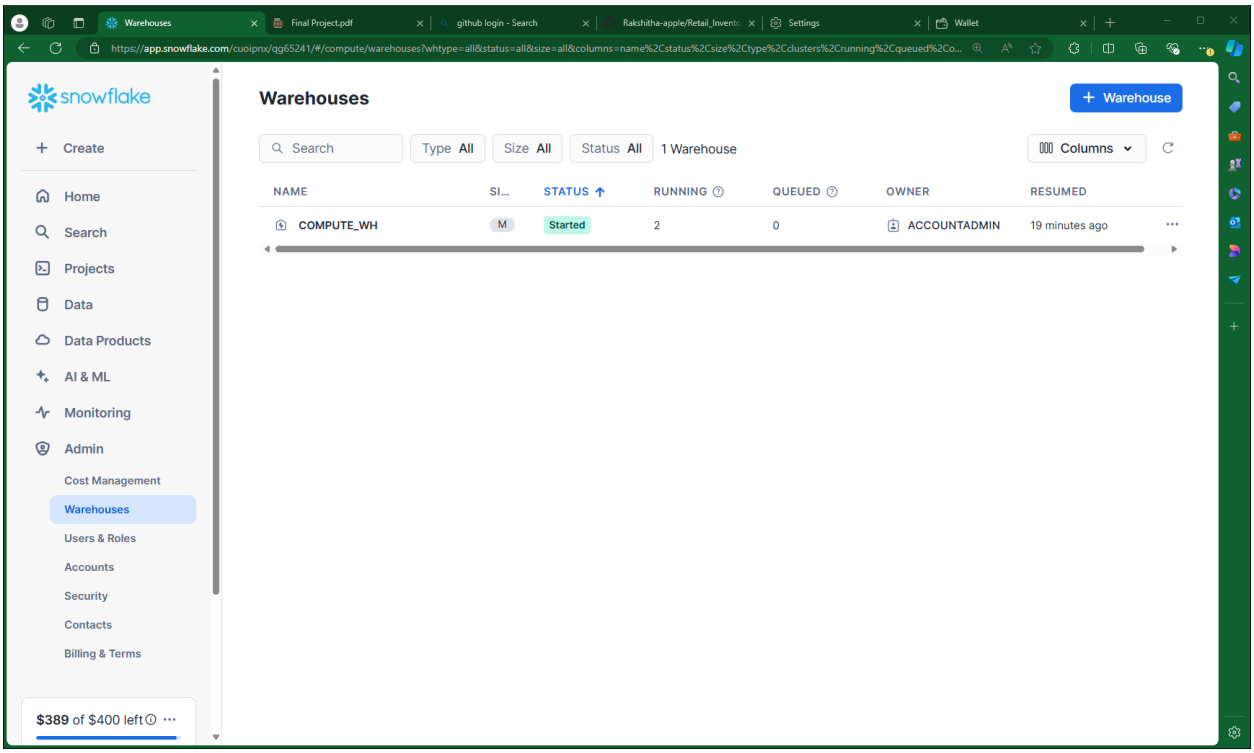
**ai16587.canada-central.azure.snowflakecomputing.com**

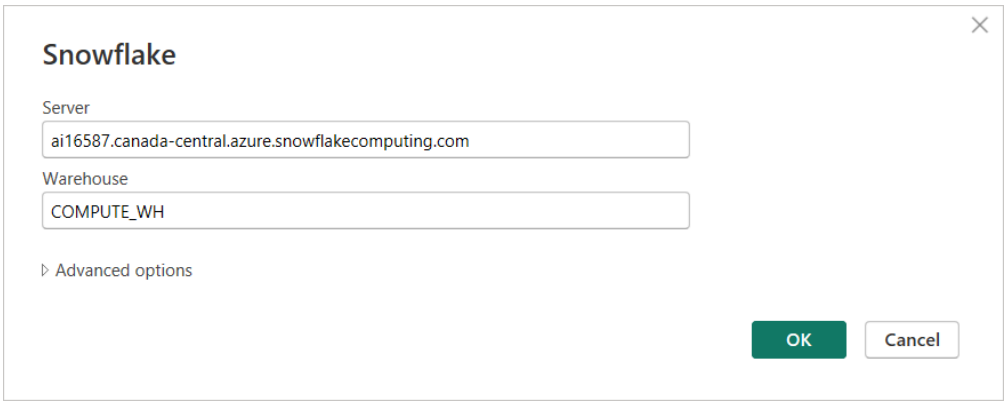


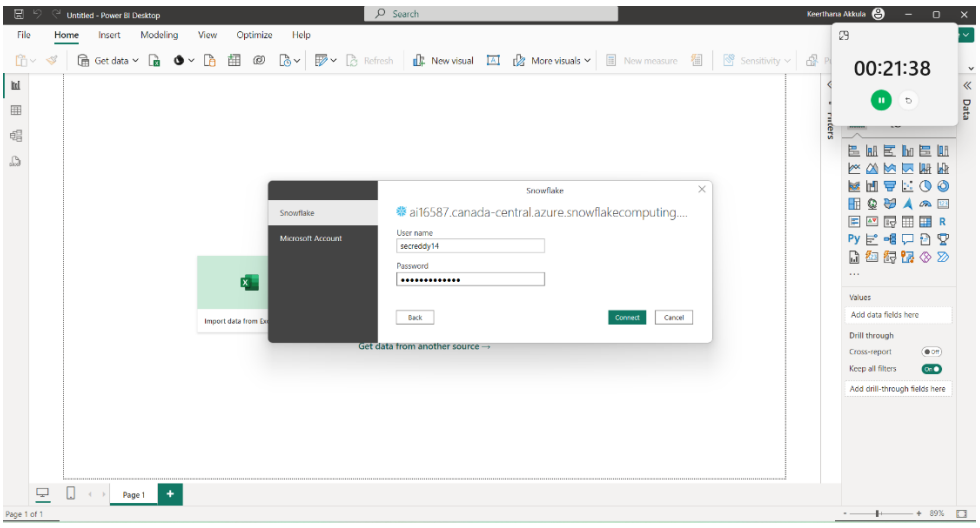




Microsoft Wallet
10,100
US$0.00
Home
Payment methods
Passwords
Personal info
Order tracking
Donation
Settings
Passwords / snowflakecomputing.com
secreddy14
Password
Site
https://ai16587.canada-central.azure.snowflakecomp...
Notes
No note added
Edit
Delete
secreddy14
Password
Site
https://cuoipnx-qg65241.snowflakecomputing.com/c...
Notes
No note added
Edit
Delete
+10
10
Level 1
Satisfied with Wallet?


Sign in to Snowflake
Username
secreddy14
Password
Forgot password
Sign in
We process your personal information according to our
Privacy Notice


Warehouses
Final Project.pdf
github login - Search
Rakshitha-apple/Retail_Invento
Settings
Wallet
https://app.snowflake.com/cuoipnx/qg65241/#/compute/warehouses?whtype=all&status=all&size=all&columns=name%2Cstatus%2Csize%2Ctype%2Cclusters%2Crunning%2Cqueued%2Co...
snowflake
Create
Home
Search
Projects
Data
Data Products
AI & ML
Monitoring
Admin
Cost Management
Warehouses
Users & Roles
Accounts
Security
Contacts
Billing & Terms
$389 of $400 left
Warehouses
+ Warehouse
Search
Type All
Size All
Status All
1 Warehouse
Columns
NAME
SI...
STATUS
RUNNING
QUEUED
OWNER
RESUMED
COMPUTE_WH
M
Started
2
0
ACCOUNTADMIN
19 minutes ago


Snowflake
Server
ai16587.canada-central.azure.snowflakecomputing.com
Warehouse
COMPUTE_WH
Advanced options
OK
Cancel


Untitled - Power BI Desktop
File
Home
Insert
Modeling
View
Optimize
Help
Get data
New visual
More visuals
00:21:38
Snowflake
ai16587.canada-central.azure.snowflakecomputing....
Microsoft Account
User name
secreddy14
Password
Back
Connect
Cancel
Get data from another source
Page 1

## Navigator

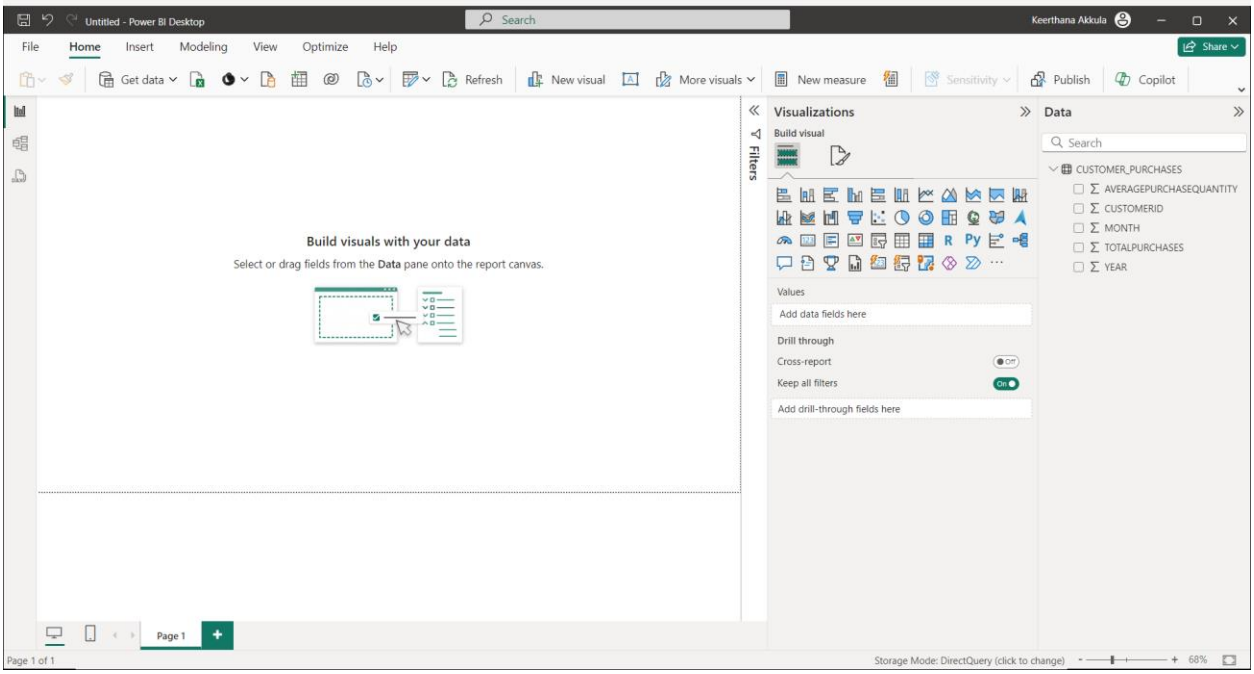
Display Options

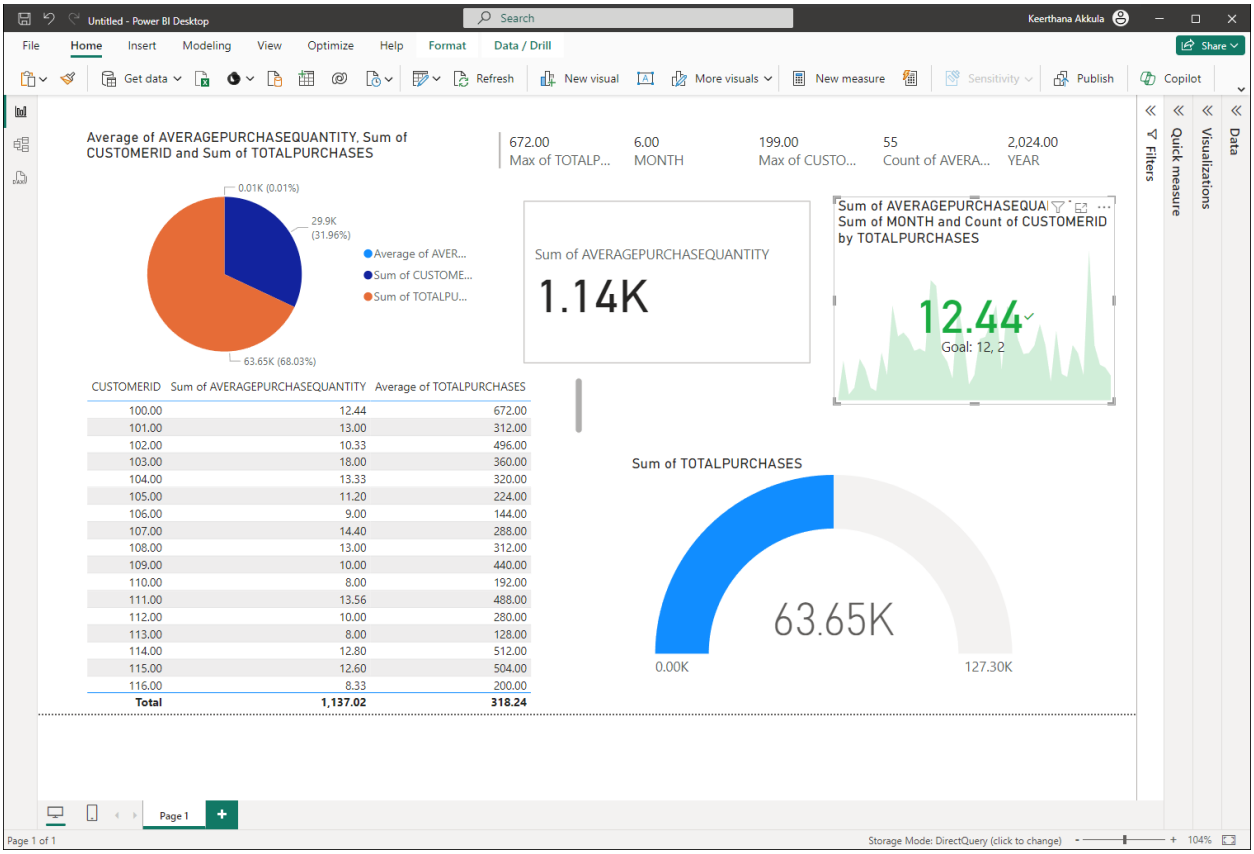
- ai16587.canada-central.azure.snowflakecompu...
  - SNOWFLAKE
  - SNOWFLAKE_SAMPLE_DATA
  - SNOWFLAKEDATABASE [2]
    - PUBLIC
    - SNOWFLAKESCHEMA [1]
      - CUSTOMER_PURCHASES

### CUSTOMER_PURCHASES

| CUSTOMERID | YEAR | MONTH | TOTALPURCHASES | AVERAGEPURCHASEQUANT... |
|---|---|---|---|---|
| 102 | 2024 | 6 | 496 | 5.1666666 |
| 146 | 2024 | 6 | 248 | 5.1666666 |
| 180 | 2024 | 6 | 280 | 4.3 |
| 111 | 2024 | 6 | 488 | 6.7777777 |
| 154 | 2024 | 6 | 112 | 4.6666666 |
| 195 | 2024 | 6 | 424 | 5.8888888 |
| 122 | 2024 | 6 | 184 | 5. |
| 131 | 2024 | 6 | 496 | 5.1666666 |
| 124 | 2024 | 6 | 120 | |
| 158 | 2024 | 6 | 384 | 6.8571428 |
| 190 | 2024 | 6 | 296 | 4.6 |
| 112 | 2024 | 6 | 280 | |
| 107 | 2024 | 6 | 288 | |
| 133 | 2024 | 6 | 72 | |
| 105 | 2024 | 6 | 224 | |
| 121 | 2024 | 6 | 184 | 5. |
| 173 | 2024 | 6 | 248 | |
| 167 | 2024 | 6 | 320 | 6.6666666 |
| 119 | 2024 | 6 | 256 | |
| 101 | 2024 | 6 | 312 | |
| 155 | 2024 | 6 | 416 | |
| 162 | 2024 | 6 | 192 | 3.4285714 |
| 170 | 2024 | 6 | 208 | |

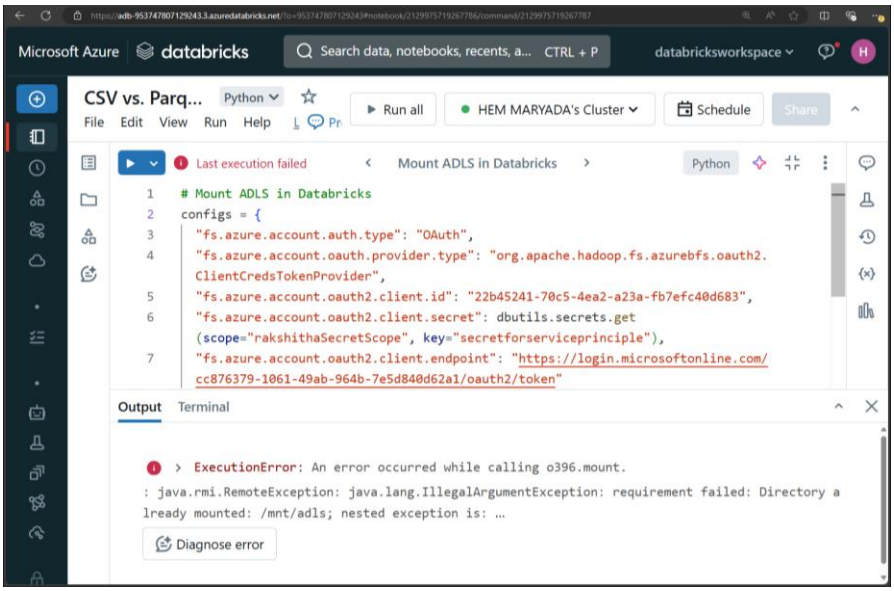
Select Related Tables | Load | Transform Data | Cancel

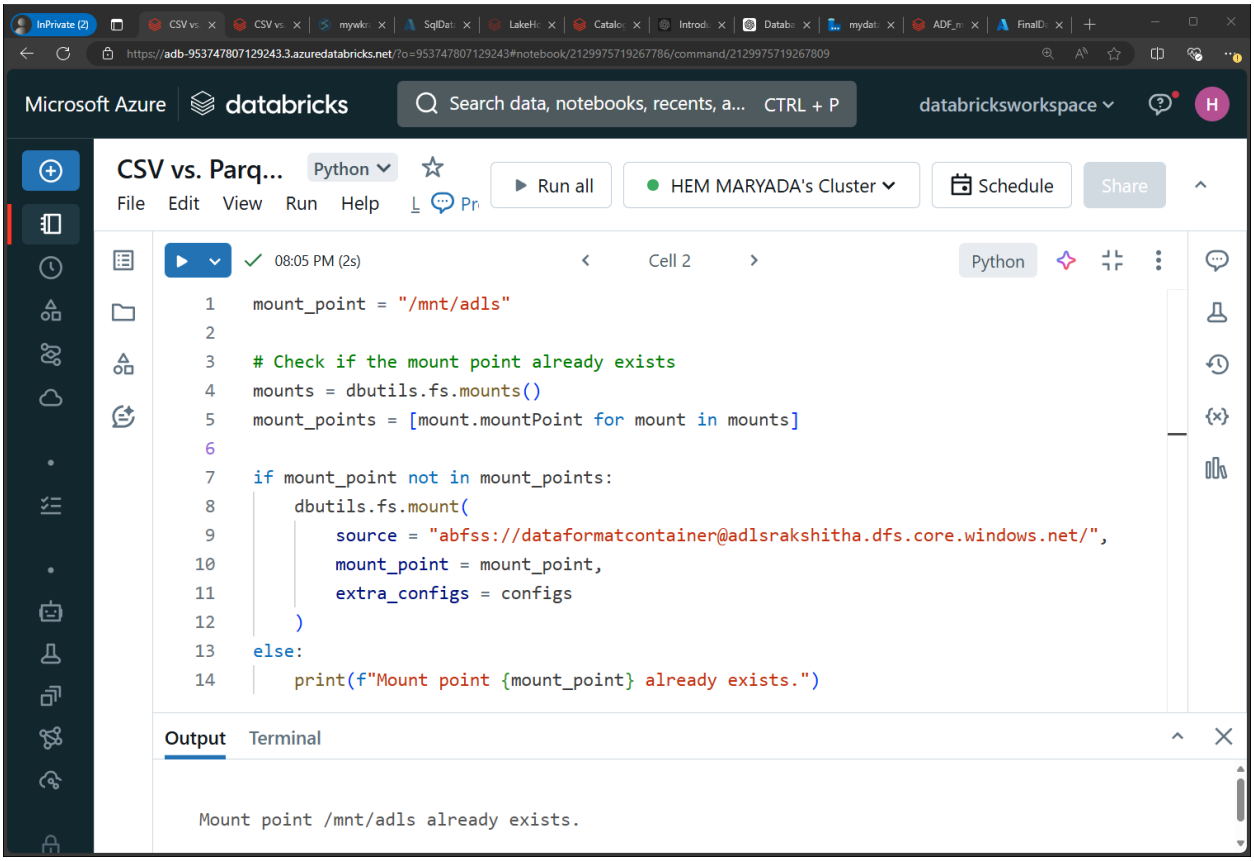


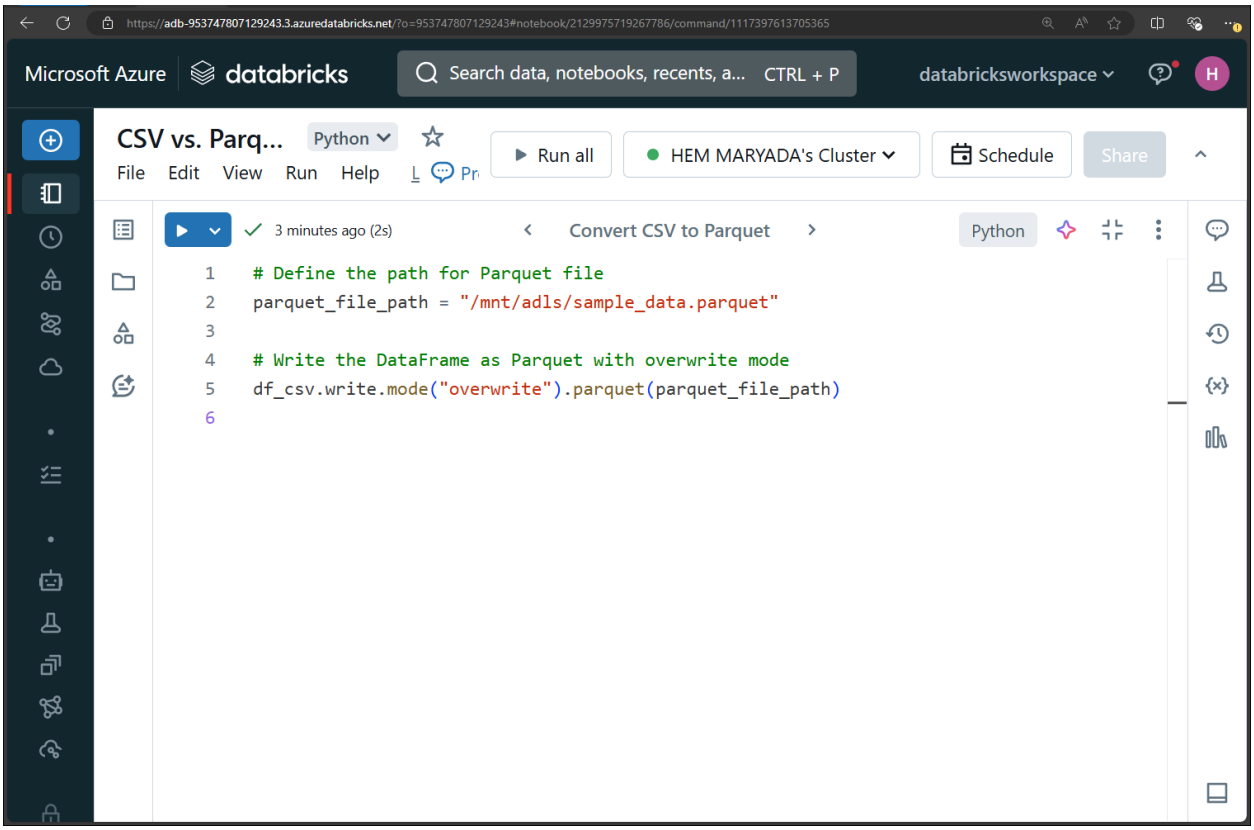


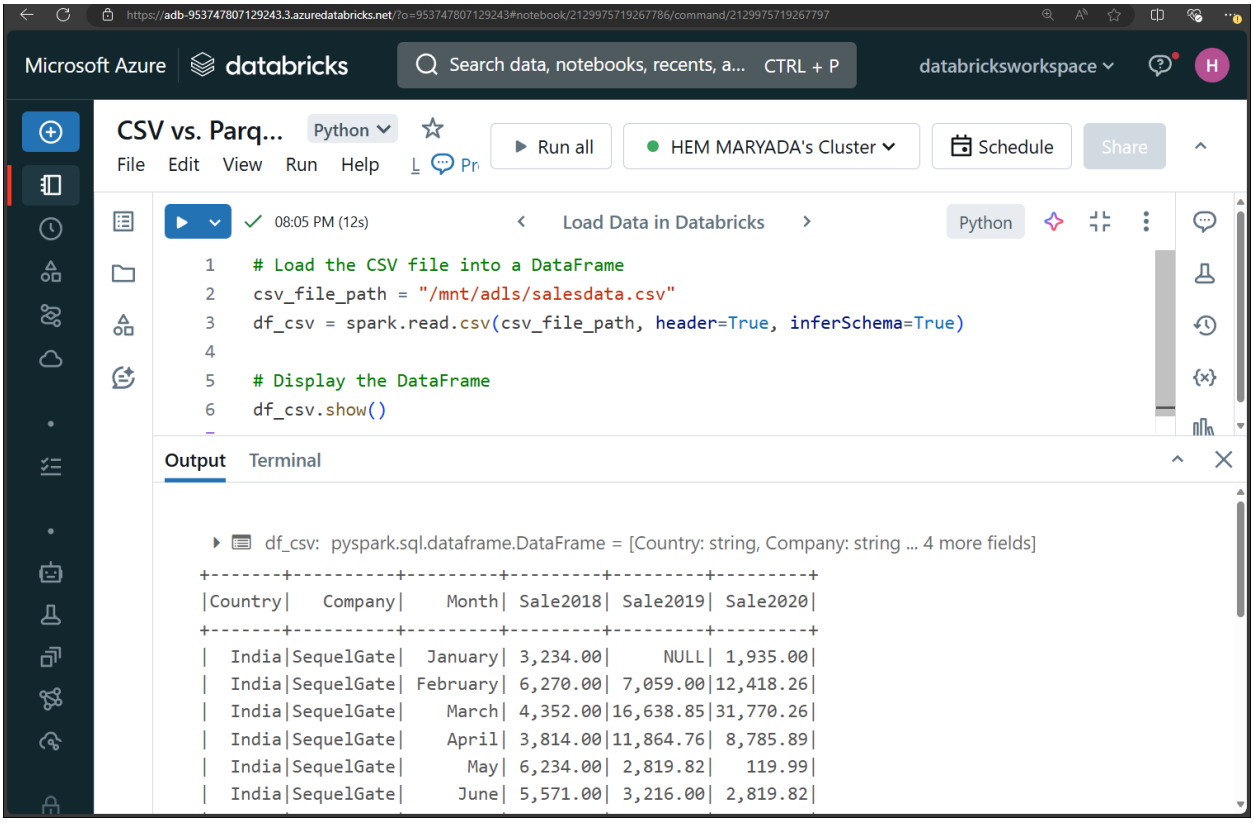

**Data Format Exploration:**

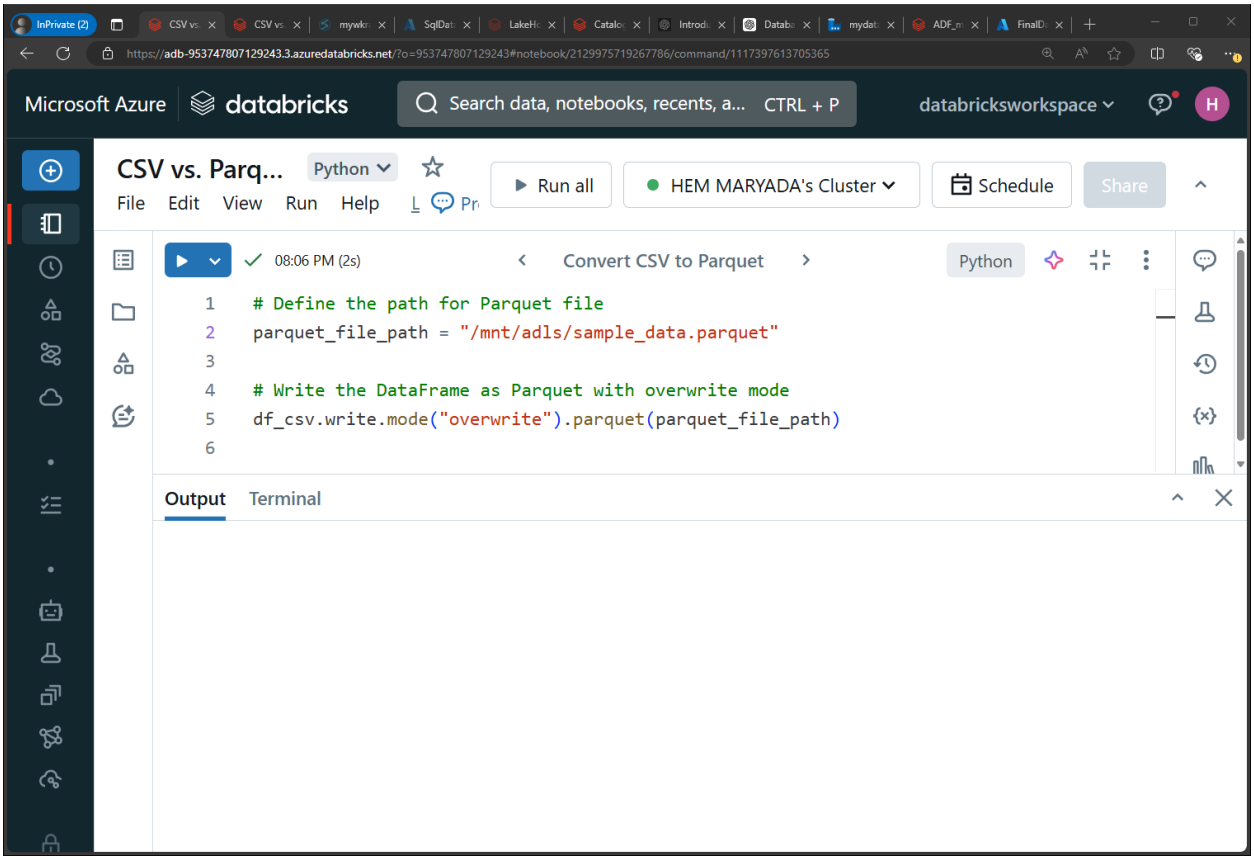
- **Experiment with different data formats (CSV vs. Parquet) and compare processing efficiency.**
- **Parquet offers better compression and faster query performance for analytical workloads.**

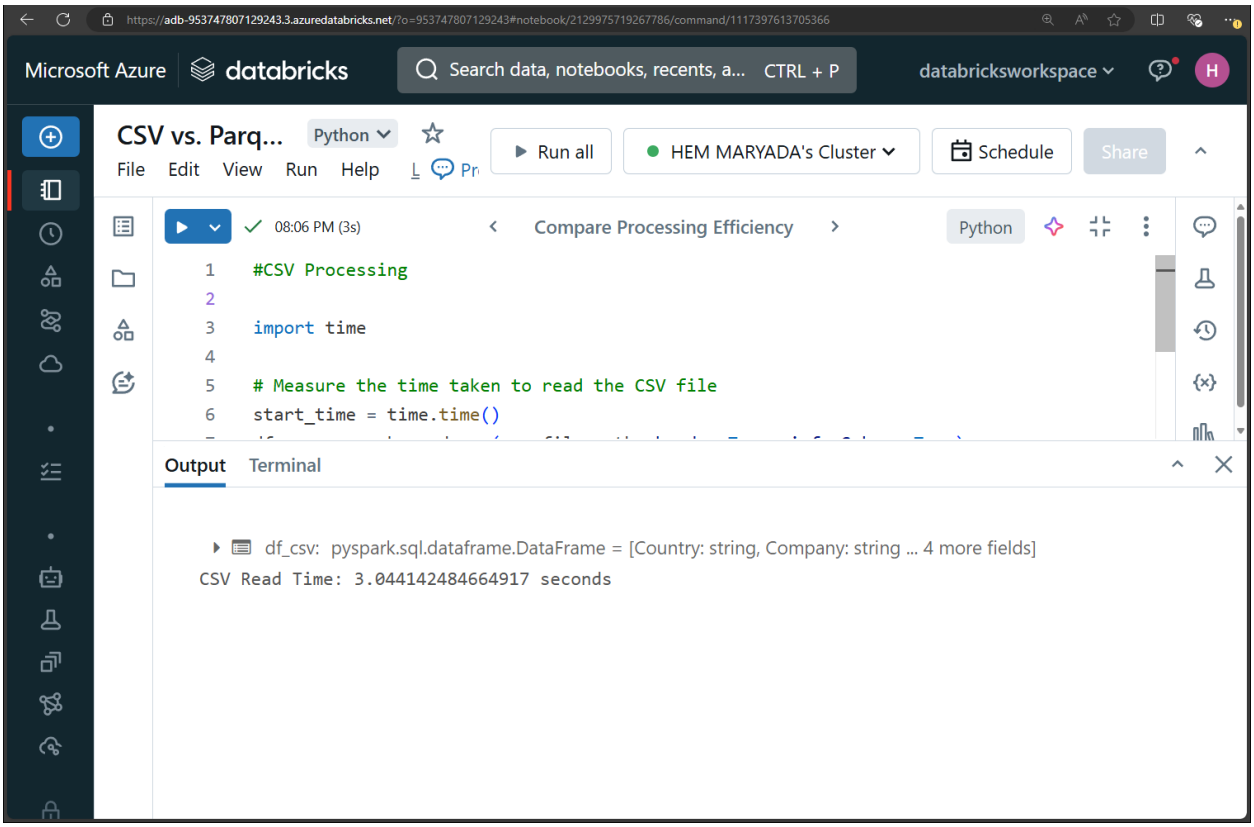


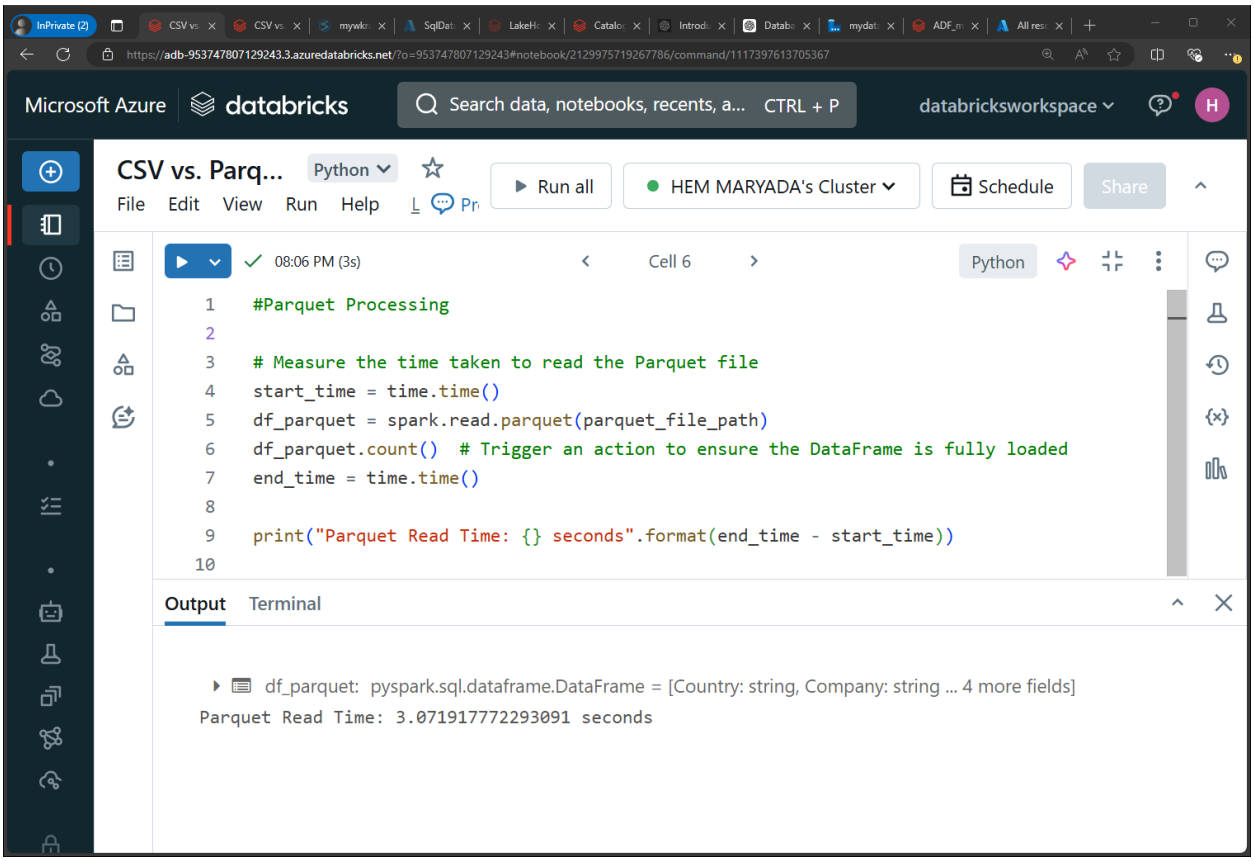

Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
08:05 PM (2s)
Cell 2
Python
1 mount_point = "/mnt/adls"
2
3 # Check if the mount point already exists
4 mounts = dbutils.fs.mounts()
5 mount_points = [mount.mountPoint for mount in mounts]
6
7 if mount_point not in mount_points:
8 dbutils.fs.mount(
9 source = "abfss://dataformatcontainer@adlsrakshitha.dfs.core.windows.net/",
10 mount_point = mount_point,
11 extra_configs = configs
12 )
13 else:
14 print(f"Mount point {mount_point} already exists.")
Output
Terminal
Mount point /mnt/adls already exists.


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
3 minutes ago (2s)
Convert CSV to Parquet
Python
1 # Define the path for Parquet file
2 parquet_file_path = "/mnt/adls/sample_data.parquet"
3
4 # Write the DataFrame as Parquet with overwrite mode
5 df_csv.write.mode("overwrite").parquet(parquet_file_path)
6


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
08:05 PM (12s)
Load Data in Databricks
Python
1 # Load the CSV file into a DataFrame
2 csv_file_path = "/mnt/adls/salesdata.csv"
3 df_csv = spark.read.csv(csv_file_path, header=True, inferSchema=True)
4
5 # Display the DataFrame
6 df_csv.show()
Output
Terminal
df_csv: pyspark.sql.dataframe.DataFrame = [Country: string, Company: string ... 4 more fields]
+-------+----------+---------+---------+---------+---------+
|Country| Company| Month| Sale2018| Sale2019| Sale2020|
+-------+----------+---------+---------+---------+---------+
| India|SequelGate| January| 3,234.00| NULL| 1,935.00|
| India|SequelGate| February| 6,270.00| 7,059.00|12,418.26|
| India|SequelGate| March| 4,352.00|16,638.85|31,770.26|
| India|SequelGate| April| 3,814.00|11,864.76| 8,785.89|
| India|SequelGate| May| 6,234.00| 2,819.82| 119.99|
| India|SequelGate| June| 5,571.00| 3,216.00| 2,819.82|


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
08:06 PM (2s)
Convert CSV to Parquet
Python
1 # Define the path for Parquet file
2 parquet_file_path = "/mnt/adls/sample_data.parquet"
3
4 # Write the DataFrame as Parquet with overwrite mode
5 df_csv.write.mode("overwrite").parquet(parquet_file_path)
6
Output
Terminal


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
08:06 PM (3s)
Compare Processing Efficiency
Python
#CSV Processing
import time
# Measure the time taken to read the CSV file
start_time = time.time()
Output
Terminal
df_csv: pyspark.sql.dataframe.DataFrame = [Country: string, Company: string ... 4 more fields]
CSV Read Time: 3.044142484664917 seconds


Microsoft Azure
databricks
Search data, notebooks, recents, a... CTRL + P
databricksworkspace
CSV vs. Parq...
Python
File Edit View Run Help
Run all
HEM MARYADA's Cluster
Schedule
Share
08:06 PM (3s)
Cell 6
Python
#Parquet Processing
# Measure the time taken to read the Parquet file
start_time = time.time()
df_parquet = spark.read.parquet(parquet_file_path)
df_parquet.count()  # Trigger an action to ensure the DataFrame is fully loaded
end_time = time.time()
print("Parquet Read Time: {} seconds".format(end_time - start_time))
Output
Terminal
df_parquet: pyspark.sql.dataframe.DataFrame = [Country: string, Company: string ... 4 more fields]
Parquet Read Time: 3.071917772293091 seconds