UNIVERSITY OF KELANIYA – SRI LANKA
FACULTY OF COMPUTING AND TECHNOLOGY

CSCI 32092 - Data Mining and Warehousing Coursework

Academic Year 2024

REPORT OF THE COURSEWORK

Student no: CS/2019/053

Student Name:M.D.H.M.Gunathilka

SECTION A: DIMENSIONAL MODELING

Do the following tasks for designing the logical data.

I. Identify contexts for measures for sales analytics and make a Bus Matrix, showing all the Possible attributes of identified dimensions.

Contexts for Measures:

1. **Sales**
2. **Marketing Expenditure**
3. **Real-Time Sales**

Dimensions Identified:

1. **Time**
2. **Customer**
3. **Product**
4. **Region**

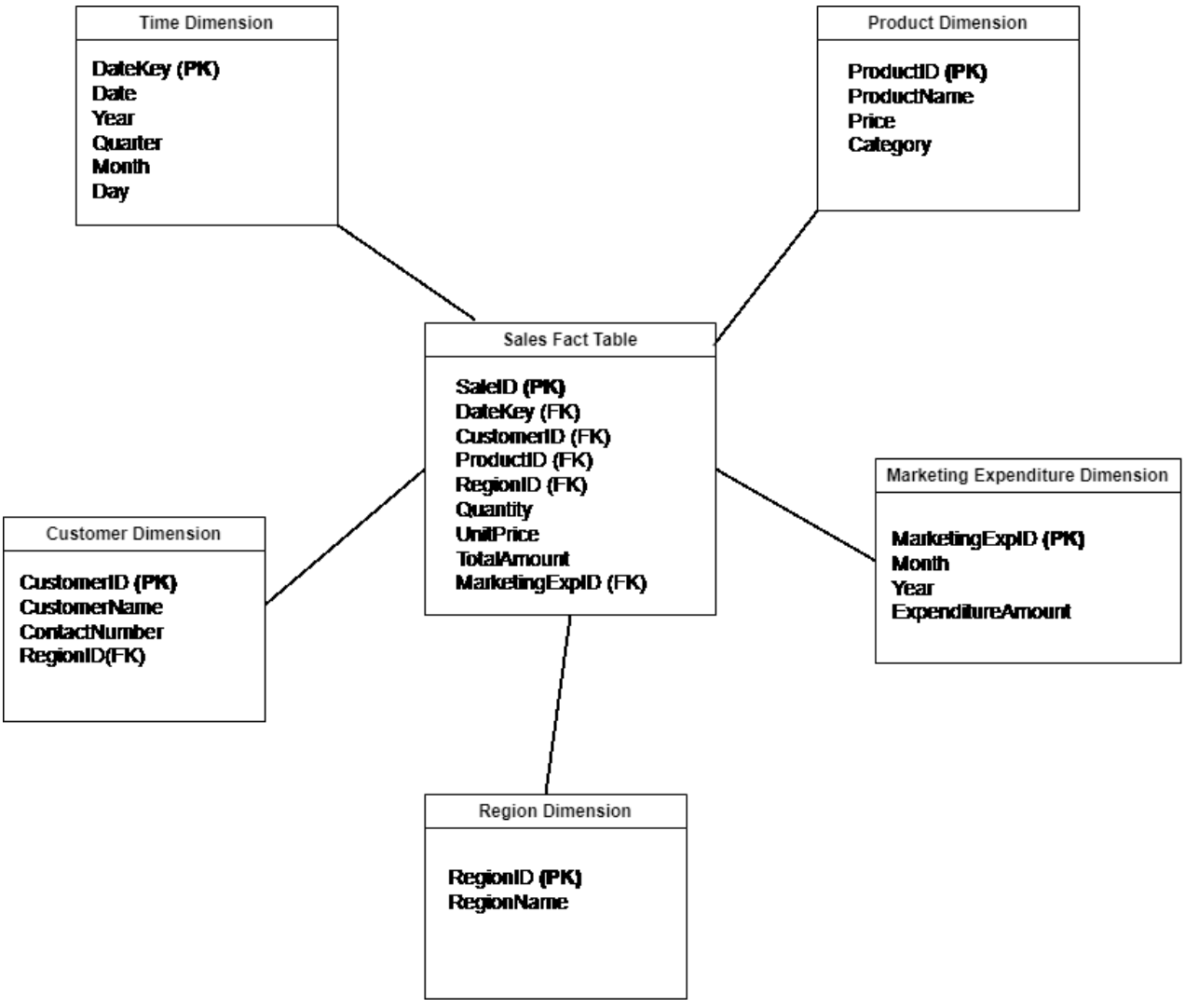
Possible Attributes of Identified Dimensions:

1. **Time Dimension:**
   - Year
   - Quarter
   - Month
   - Day
2. **Customer Dimension:**
   - CustomerID
   - CustomerName
   - ContactNumber
   - RegionID
3. **Product Dimension:**
   - ProductID
   - ProductName
   - Price
   - Category
4. **Region Dimension:**
   - RegionID
   - RegionName

Bus Matrix

Measure	Time	Customer	Product	Region
Total Sales Amount	✓	✓	✓	✓
Total Quantity Sold	✓	✓	✓	✓
Marketing Expenditure	✓			
Unit Price			✓	
Sales Count	✓	✓	✓	✓

II. Design a star schema for the above scenario. Make sure to consider data coming from all the **3 data sources** when you are deciding 'fields' in your dimensions and fact table



SECTION B: PHYSICAL DW DESIGN AND ETL USING ADF

Assumptions for Completing the POC Using Azure Services

1. Source Database Setup:
- The source database is created using Azure SQL Database according to the provided ER diagram.
- The table structures is created exactly as per the given ER diagram, with appropriate primary and foreign key constraints.

2. Data Insertion:
- The data will be consistent with the defined data types and relationships in the ER diagram.

3. Primary Key and Foreign Key Constraints:
- Each table will have primary keys as defined: `SaleID` for Sales, `CustomerID` for Customers, `ProductID` for Products, and `RegionID` for Regions.
- Foreign key constraints will be enforced to maintain data integrity between tables.

4. Data Types and Sizes:
- Assumptions about the data types and sizes for the table columns will be made if not explicitly specified. For example, `CustomerName` might be assumed as `VARCHAR(100)`, `ContactNumber` as `VARCHAR(15)`, `RegionName` as `VARCHAR(50)`, etc.

5. Date Handling:
- The `Date` column in the Sales table will be assumed to be in the format `YYYY-MM-DD` and will be of data type `DATE`.

6. Azure Data Factory (ADF) Configuration:
- Azure Data Factory will be used to create and configure data flows for ETL (Extract, Transform, Load) processes.
- Linked services and datasets will be configured to connect to the Azure SQL Database.

7. Lookup Transformations in ADF:
- Lookup transformations in ADF will be used to handle foreign key relationships and retrieve dimension keys.
- Multiple lookup transformations can be used within the same data flow to enrich the Sales fact table with dimension keys from the related tables.

8. ETL Process:
- The ETL process will involve extracting data from the source tables, performing necessary transformations (e.g., extracting year and month from the `Date` column), and loading the data into the target fact table and dimension tables.

- Error handling and logging mechanisms will be implemented to monitor the ETL process.

9. Scalability and Performance:

- Assumptions about the scalability and performance of the Azure SQL Database and ADF will be considered, ensuring that the setup can handle the expected volume of data and ETL workload.

10. Security and Access:

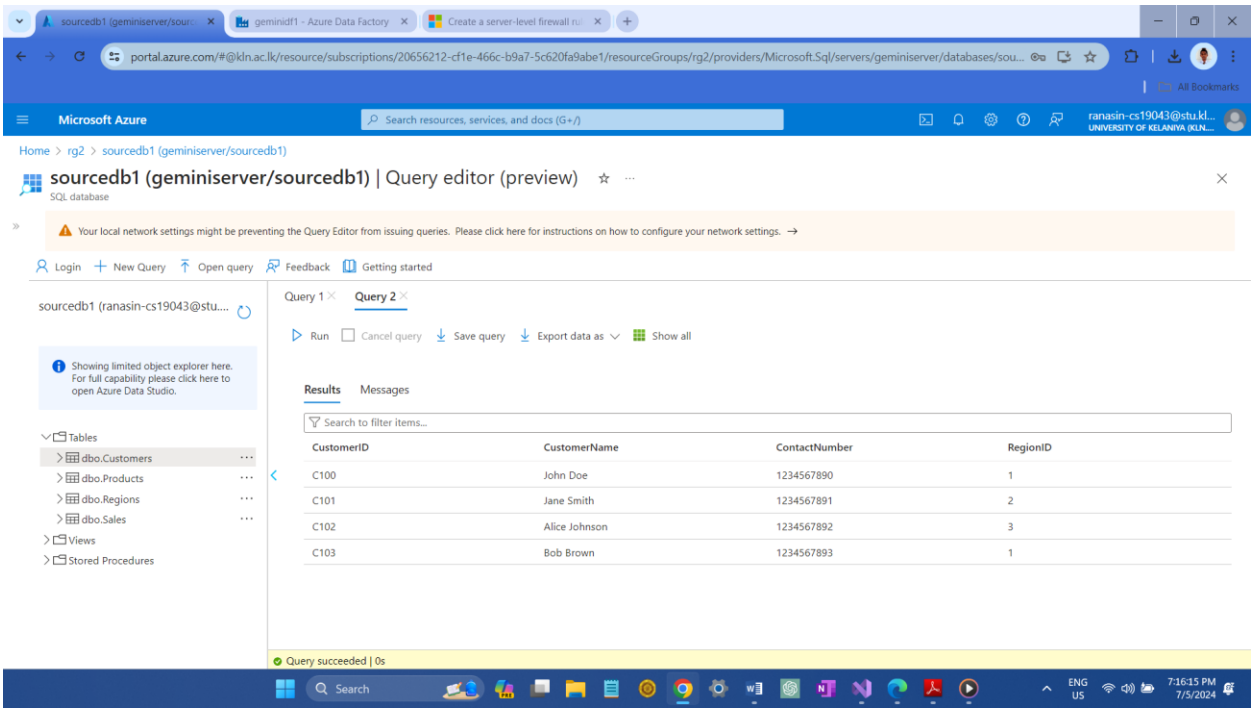
- Proper security measures, such as firewall rules and access controls, will be configured to secure the Azure SQL Database and ADF.
- The user executing the ETL processes will have the necessary permissions to read from the source tables and write to the target tables.

11. Testing and Validation:

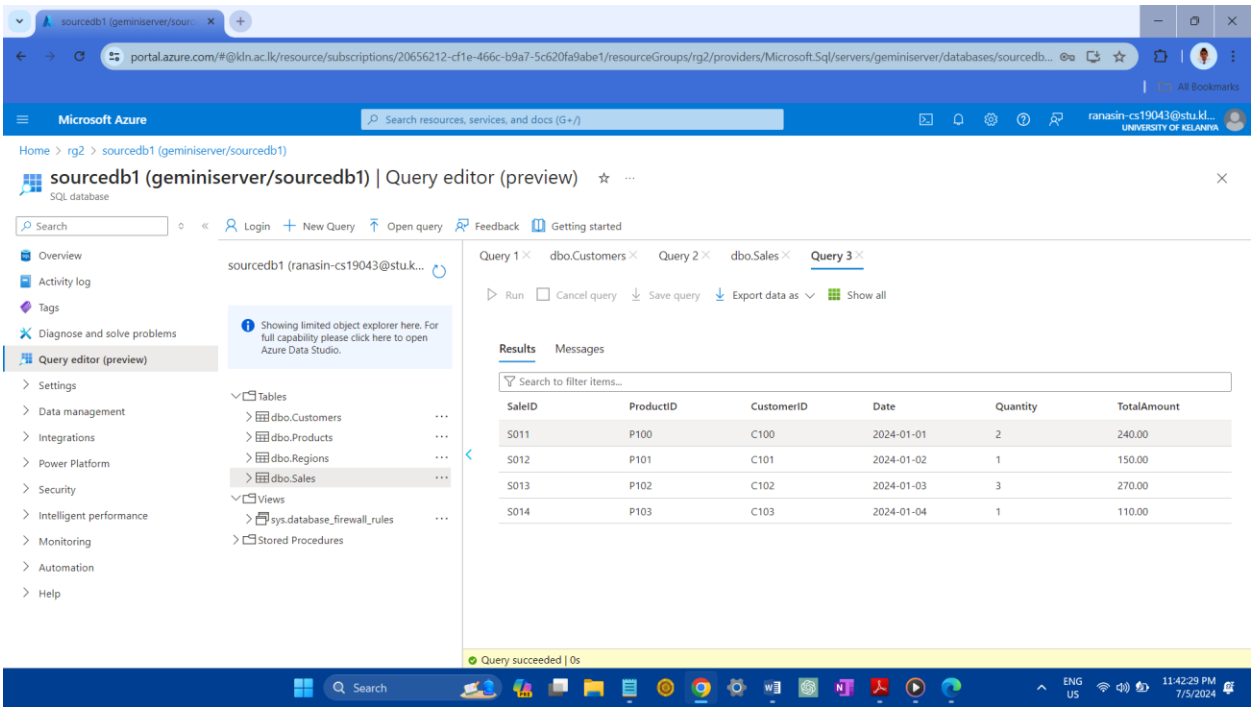
- After setting up the source database and ETL processes, thorough testing will be conducted to validate the data integrity and correctness of the transformations.
- Sample queries will be run to verify the relationships and data in the populated tables.

Tables in Source Database

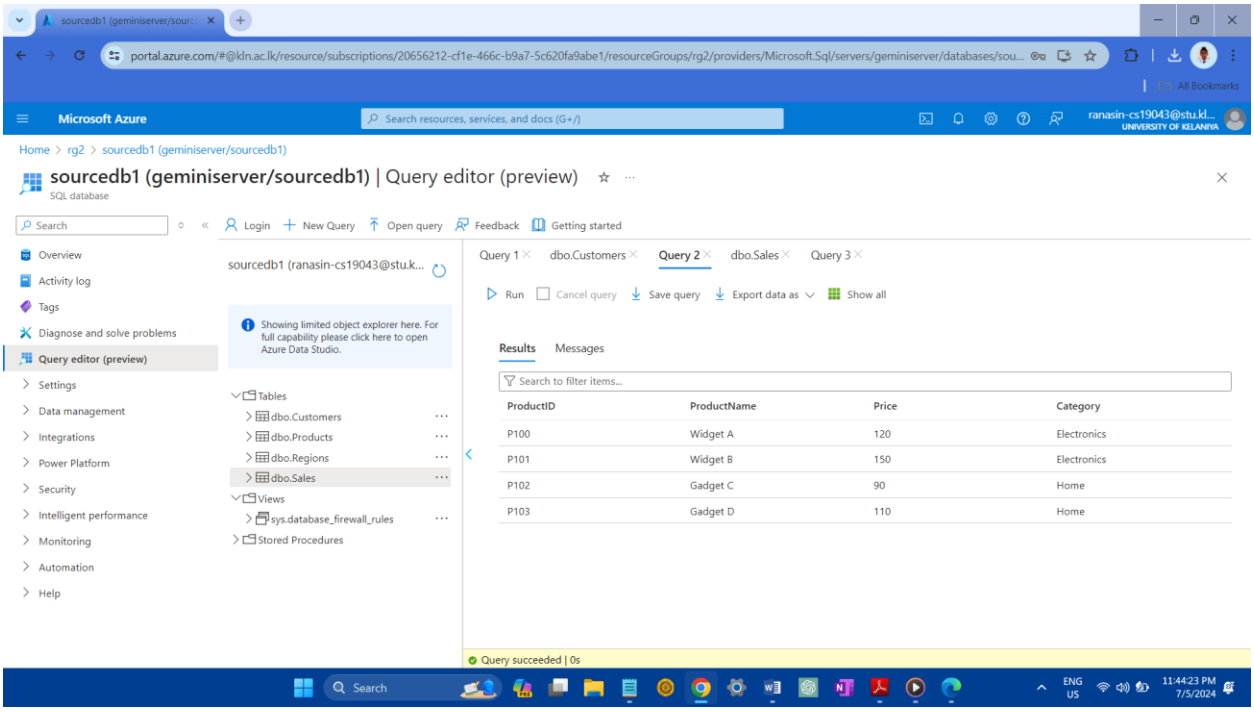
Customers Table



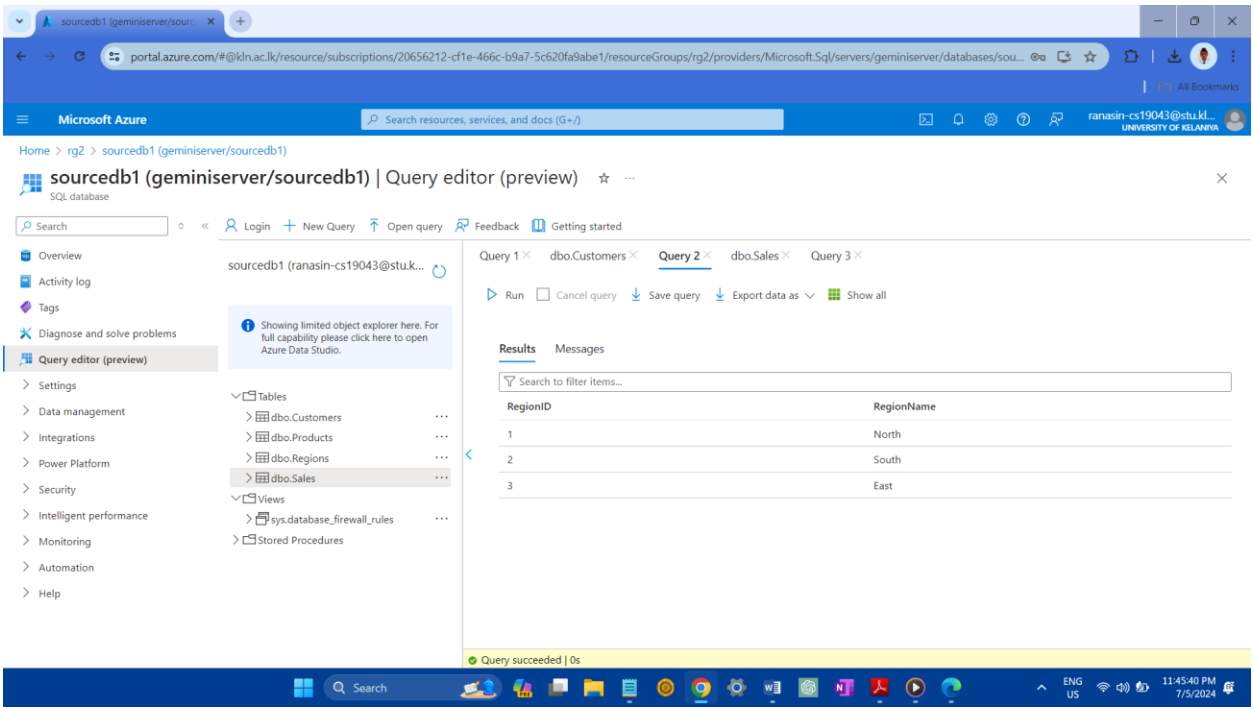
Sales Table

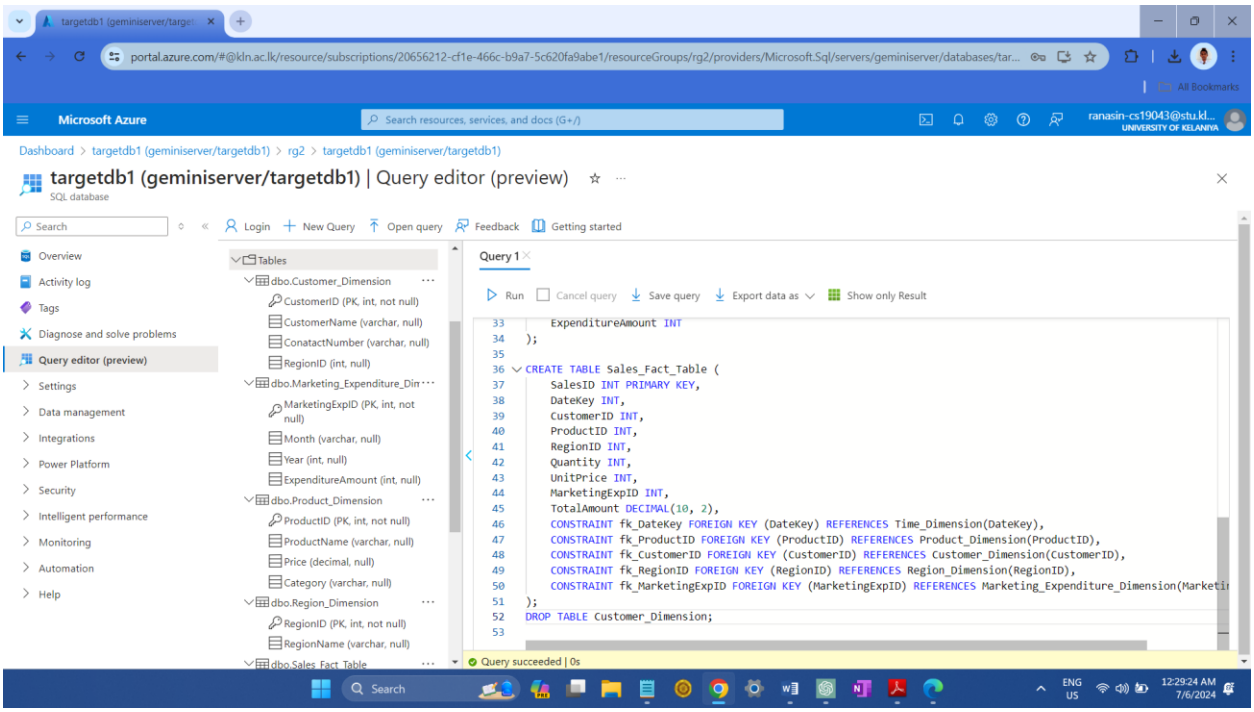


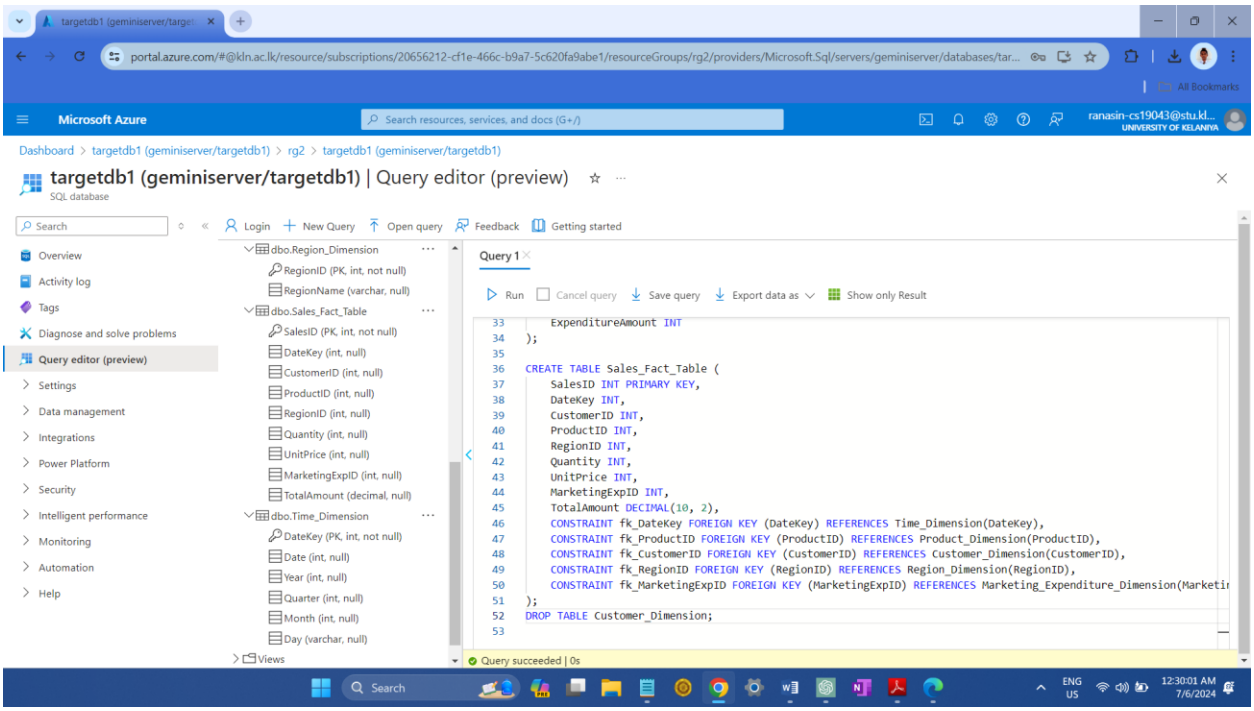
Products Table

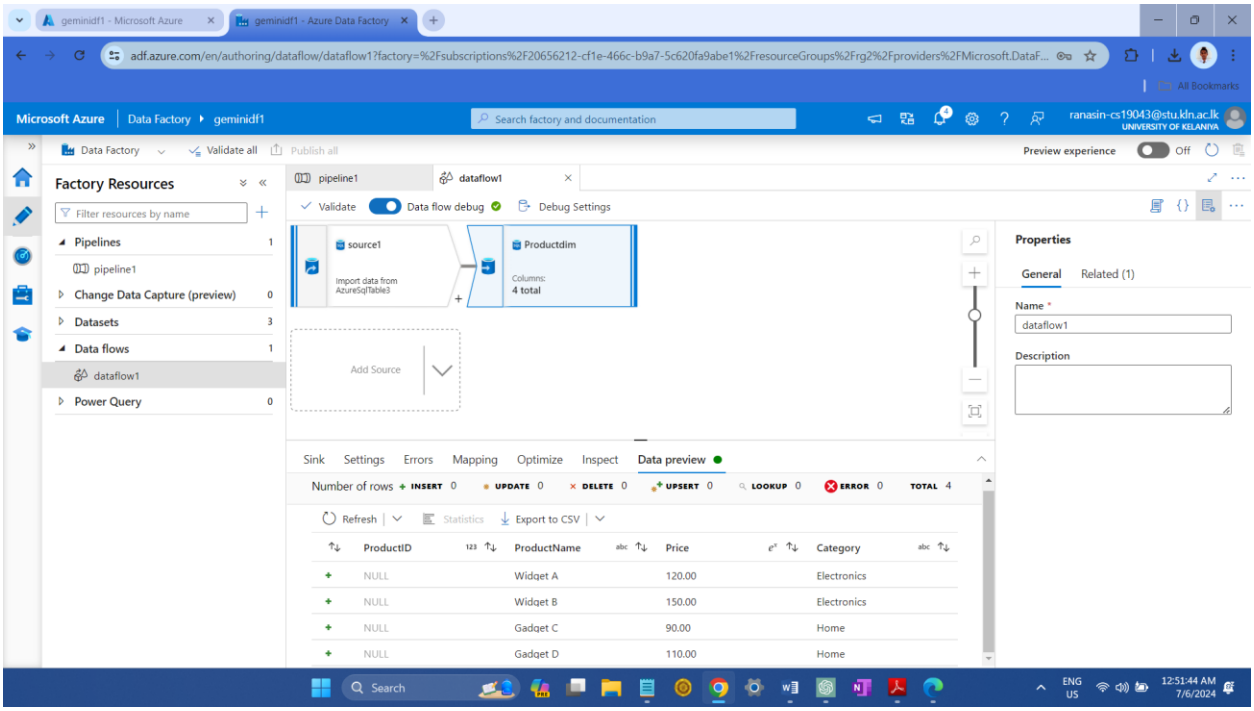


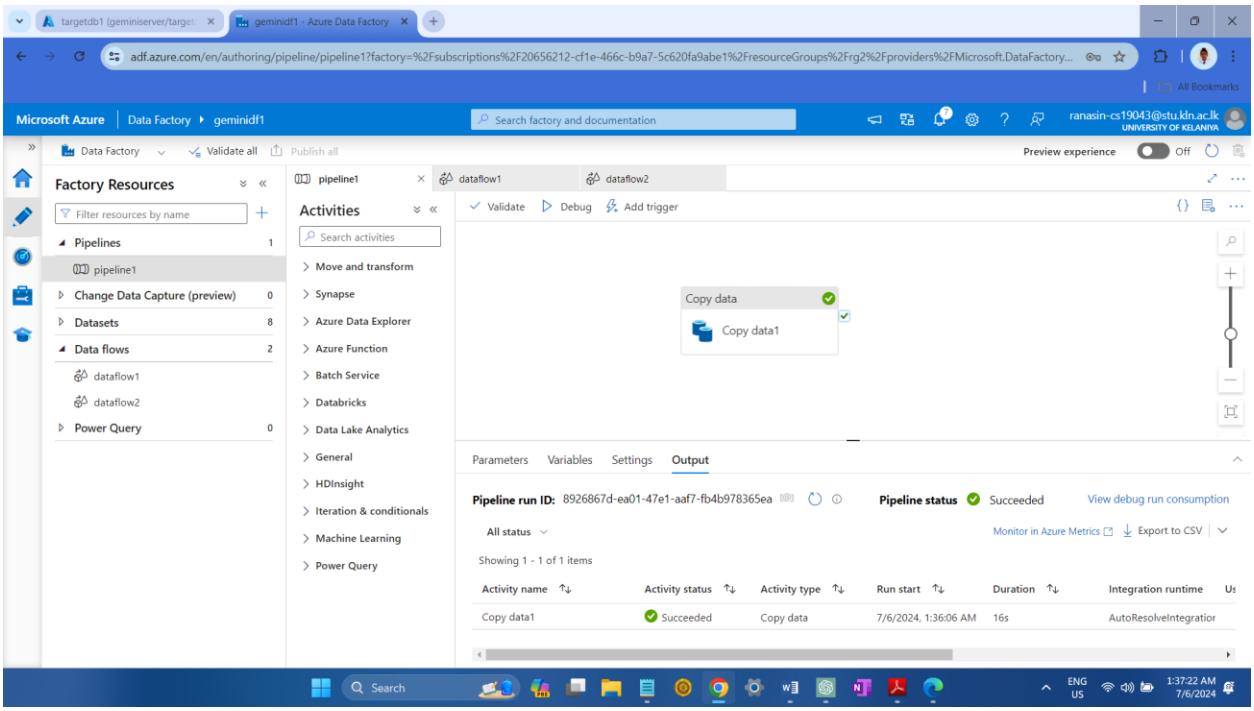
Regions Table

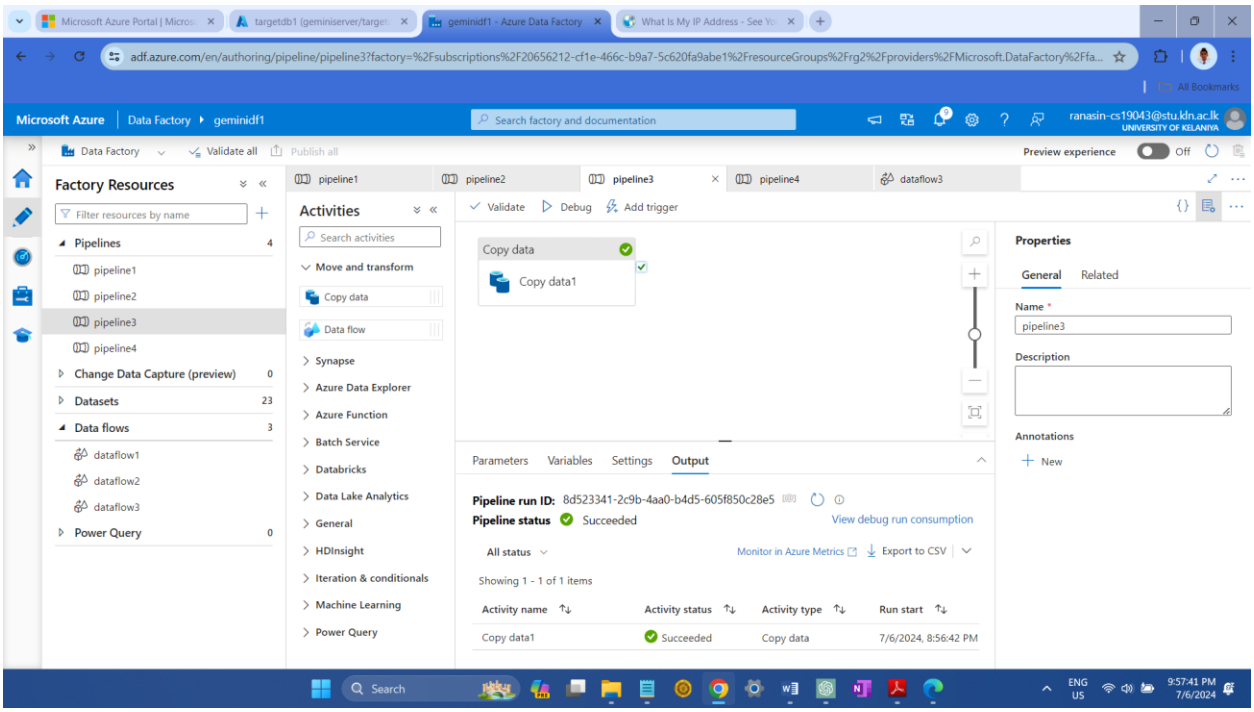

sourcedb1 (geminiserver/sourcedb1) | Query editor (preview)
SQL database
Microsoft Azure
Search resources, services, and docs (G+/)
ranasin-cs19043@stu.kl...
UNIVERSITY OF KELANIYA
Home > rg2 > sourcedb1 (geminiserver/sourcedb1)
Search
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
sourcedb1 (ranasin-cs19043@stu.k...
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.Customers
dbo.Products
dbo.Regions
dbo.Sales
Views
sys.database_firewall_rules
Stored Procedures
Query 1
dbo.Customers
Query 2
dbo.Sales
Query 3
Run
Cancel query
Save query
Export data as
Show all
Results
Messages
Search to filter items...
RegionID
RegionName
1
North
2
South
3
East
Query succeeded | 0s
Search
ENG
US
11:45:40 PM
7/5/2024

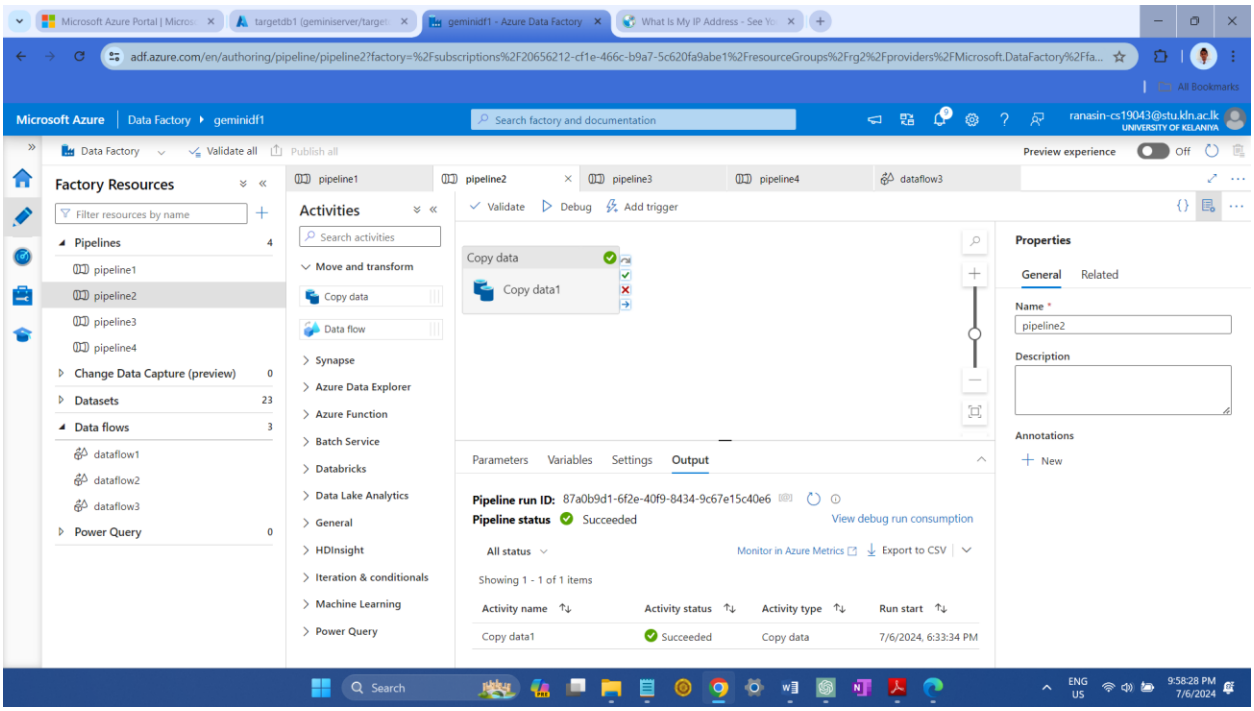

targetdb1 (geminiserver/targetdb1) | Query editor (preview)
SQL database
Microsoft Azure
Search resources, services, and docs (G+/)
ranasin-cs19043@stu.kl...
UNIVERSITY OF KELANIYA
Dashboard > targetdb1 (geminiserver/targetdb1) > rg2 > targetdb1 (geminiserver/targetdb1)
Search
Login
New Query
Open query
Feedback
Getting started
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
Tables
dbo.Customer_Dimension
CustomerID (PK, int, not null)
CustomerName (varchar, null)
ConatactNumber (varchar, null)
RegionID (int, null)
dbo.Marketing_Expenditure_Dim
MarketingExpID (PK, int, not null)
Month (varchar, null)
Year (int, null)
ExpenditureAmount (int, null)
dbo.Product_Dimension
ProductID (PK, int, not null)
ProductName (varchar, null)
Price (decimal, null)
Category (varchar, null)
dbo.Region_Dimension
RegionID (PK, int, not null)
RegionName (varchar, null)
dbo.Sales_Fact_Table
Query 1
Run
Cancel query
Save query
Export data as
Show only Result
ExpenditureAmount INT
);
CREATE TABLE Sales_Fact_Table (
SalesID INT PRIMARY KEY,
DateKey INT,
CustomerID INT,
ProductID INT,
RegionID INT,
Quantity INT,
UnitPrice INT,
MarketingExpID INT,
TotalAmount DECIMAL(10, 2),
CONSTRAINT fk_DateKey FOREIGN KEY (DateKey) REFERENCES Time_Dimension(DateKey),
CONSTRAINT fk_ProductID FOREIGN KEY (ProductID) REFERENCES Product_Dimension(ProductID),
CONSTRAINT fk_CustomerID FOREIGN KEY (CustomerID) REFERENCES Customer_Dimension(CustomerID),
CONSTRAINT fk_RegionID FOREIGN KEY (RegionID) REFERENCES Region_Dimension(RegionID),
CONSTRAINT fk_MarketingExpID FOREIGN KEY (MarketingExpID) REFERENCES Marketing_Expenditure_Dimension(Marketi
);
DROP TABLE Customer_Dimension;
Query succeeded | 0s
Search
ENG
US
12:29:24 AM
7/6/2024

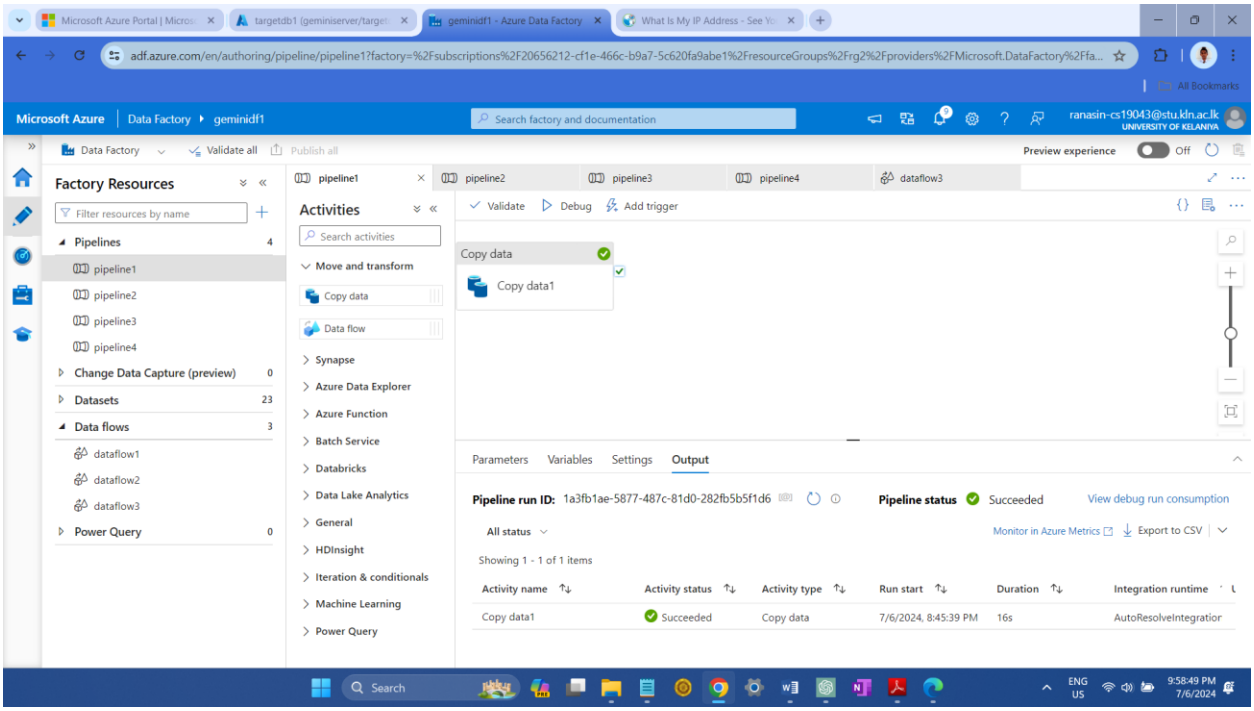

targetdb1 (geminiserver/targetdb1) | Query editor (preview)
SQL database
Query 1
Run
Cancel query
Save query
Export data as
Show only Result
Query succeeded | 0s


Factory Resources
pipeline1
dataflow1
source1
Productdim
Add Source
Properties
General
Related (1)
Name
dataflow1
Description
Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Number of rows
Refresh
Export to CSV
ProductID
ProductName
Price
Category
NULL
Widget A
120.00
Electronics
NULL
Widget B
150.00
Electronics
NULL
Gadget C
90.00
Home
NULL
Gadget D
110.00
Home

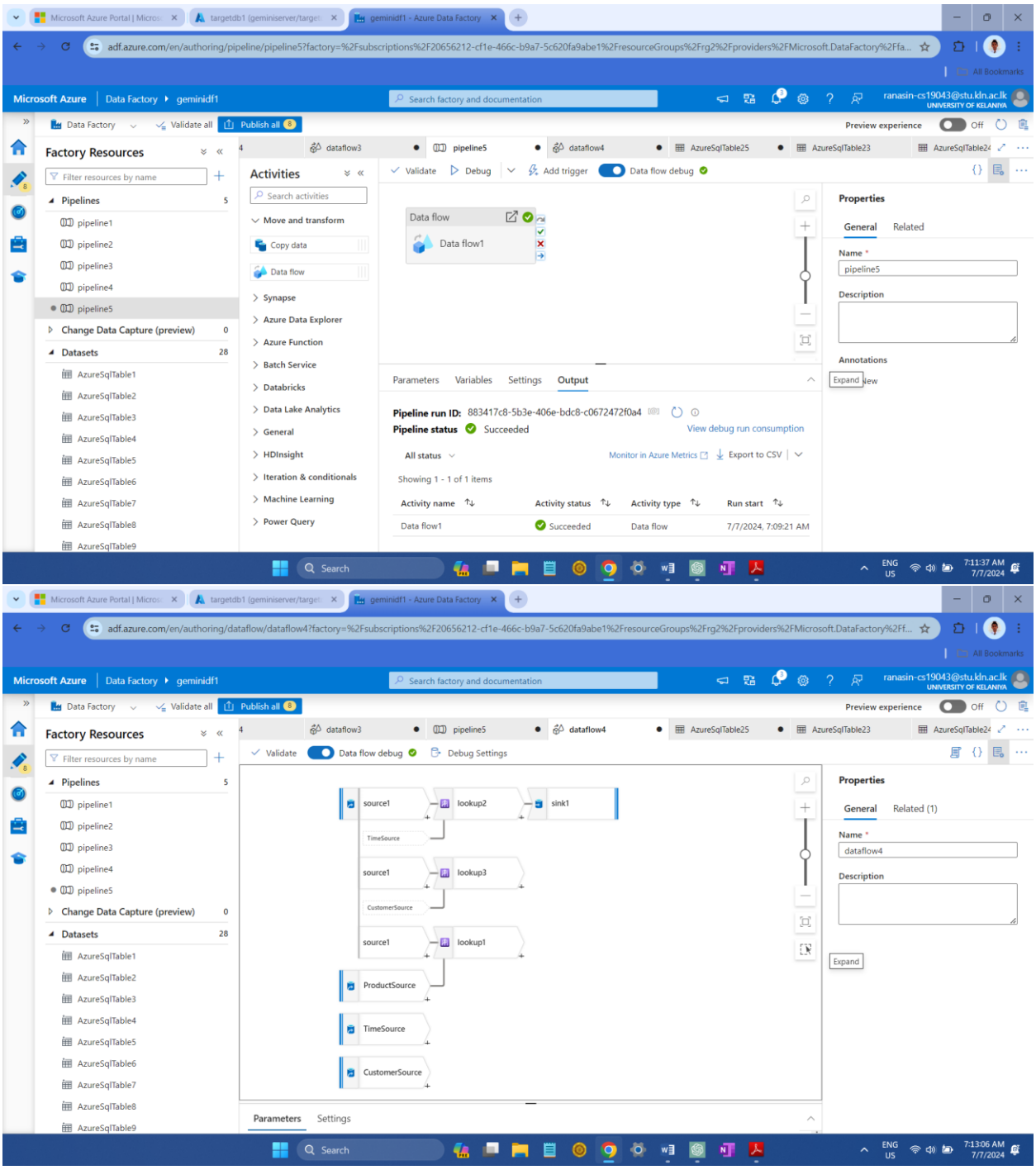

Microsoft Azure
Data Factory
geminidf1
Factory Resources
Pipelines
pipeline1
Change Data Capture (preview)
Datasets
Data flows
dataflow1
dataflow2
Power Query
Activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Copy data
Copy data1
Parameters
Variables
Settings
Output
Pipeline run ID: 8926867d-ea01-47e1-aaf7-fb4b978365ea
Pipeline status Succeeded
View debug run consumption
Activity name
Activity status
Activity type
Run start
Duration
Integration runtime
Copy data1
Succeeded
Copy data
7/6/2024, 1:36:06 AM
16s
AutoResolveIntegratior


Microsoft Azure
Data Factory
geminidf1
Factory Resources
Pipelines
pipeline1
pipeline2
pipeline3
pipeline4
Change Data Capture (preview)
Datasets
Data flows
dataflow1
dataflow2
dataflow3
Power Query
Activities
Move and transform
Copy data
Data flow
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Properties
General
Related
Name
pipeline3
Description
Annotations
New
Parameters
Variables
Settings
Output
Pipeline run ID: 8d523341-2c9b-4aa0-b4d5-605f850c28e5
Pipeline status Succeeded
View debug run consumption
Activity name
Activity status
Activity type
Run start
Copy data1
Succeeded
Copy data
7/6/2024, 8:56:42 PM

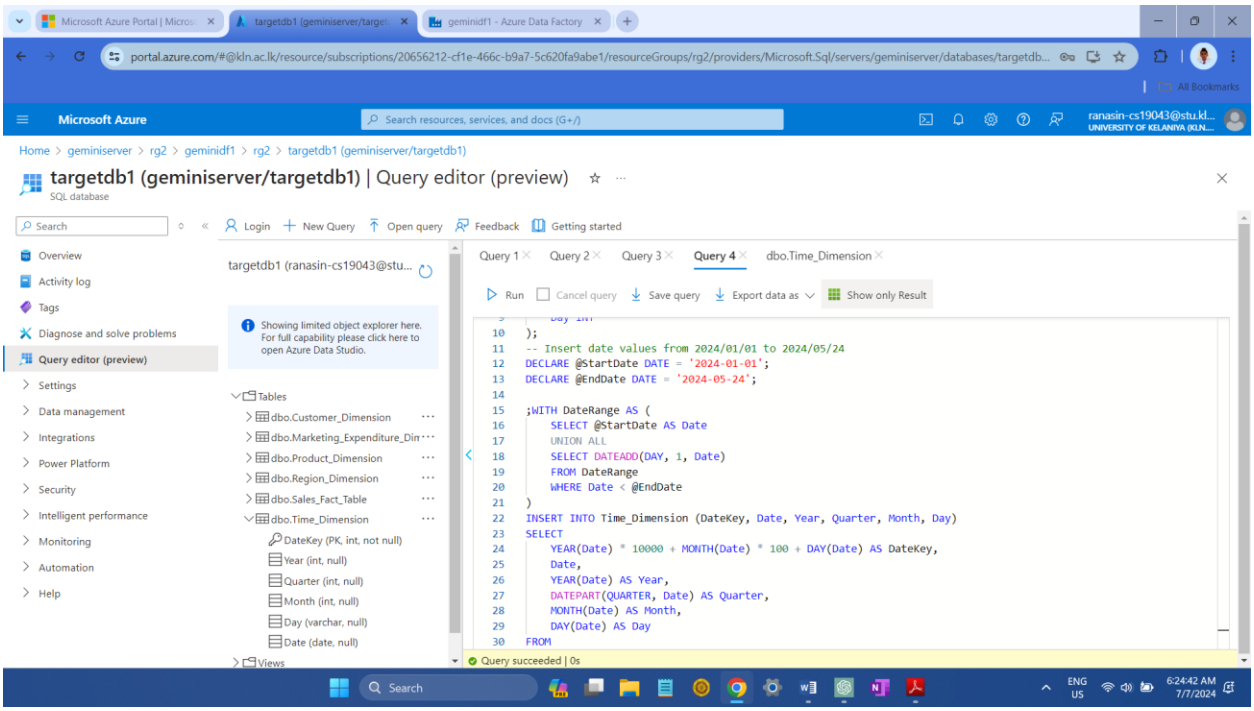


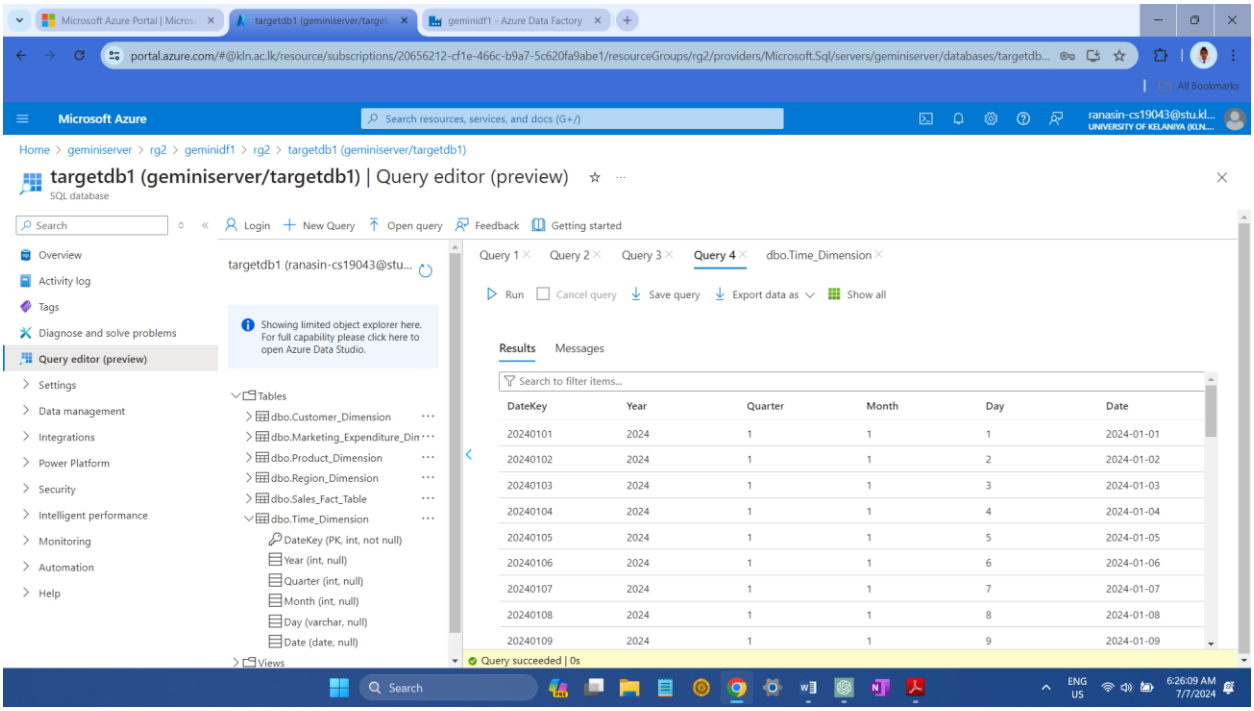


V. Use Azure Data Factory to perform ETL from 3 data sources to the target data warehouse. Use necessary schema modifiers (i.e., derived column (to calculate the total sales by unit price x quan7ty), lookup, join (to join the real-7me sales data in JSON with the historical sales data in Azure target SQL database), union, etc.)

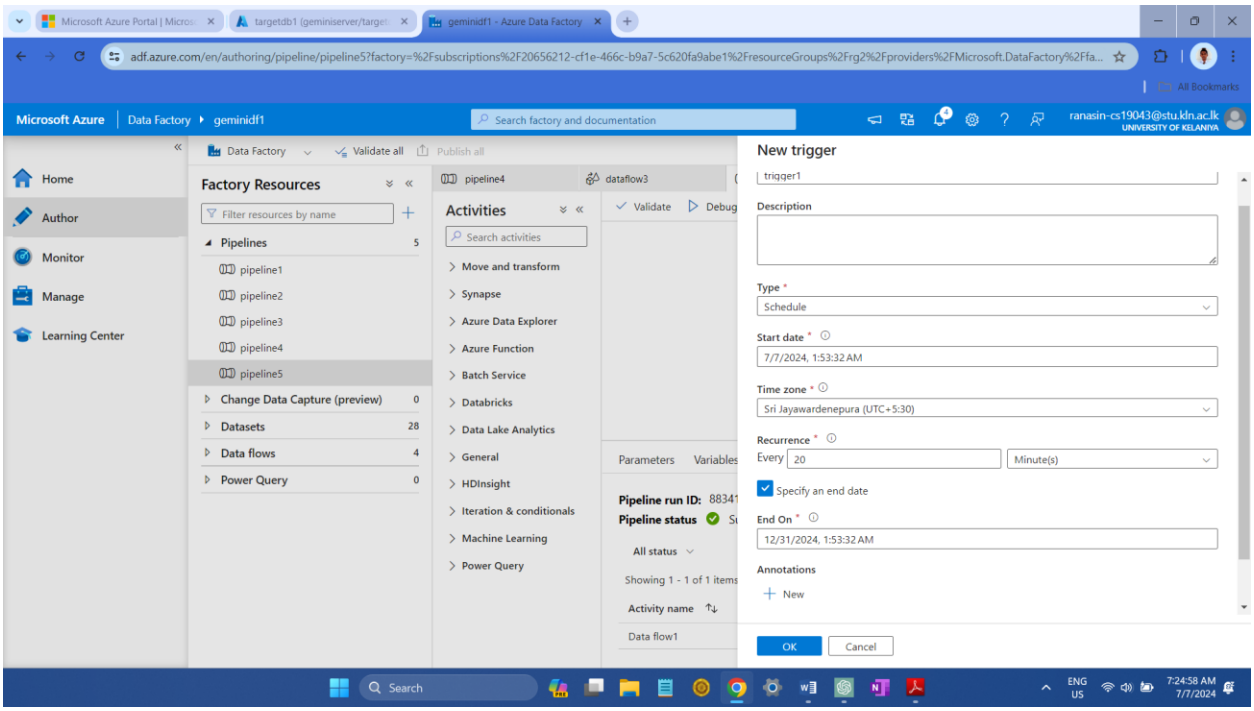


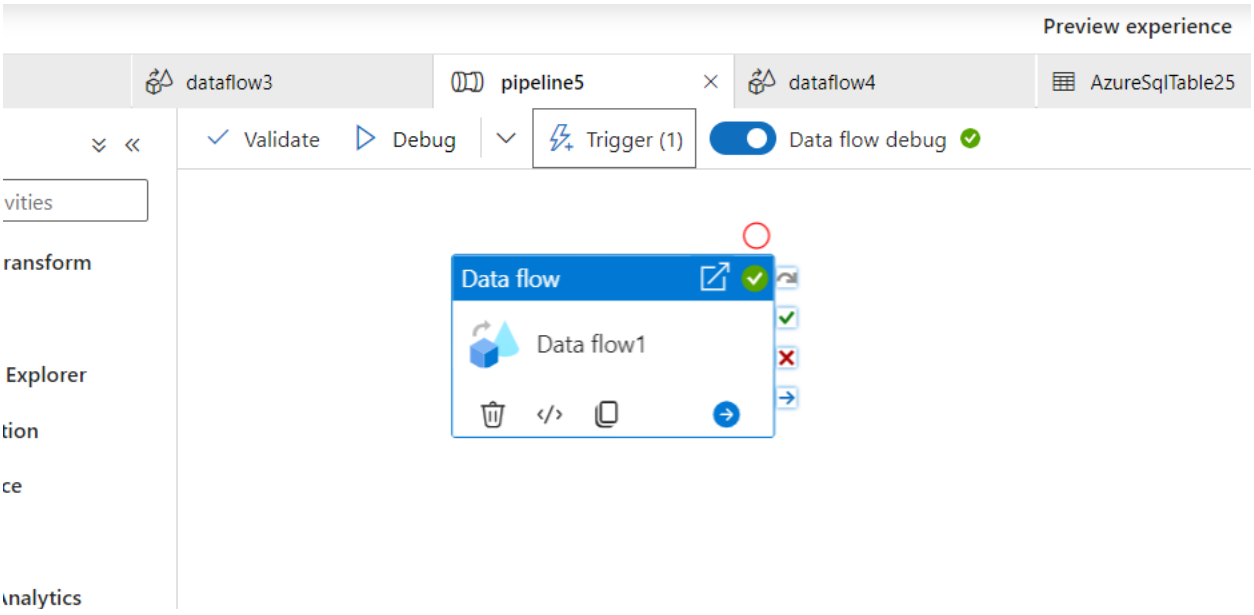
VI. Assuming analysis needs to be done in the period 2024/1/1 – 2024/5/24, populate your date dimension manually using SQL query or by performing ETL using a source CSV file.





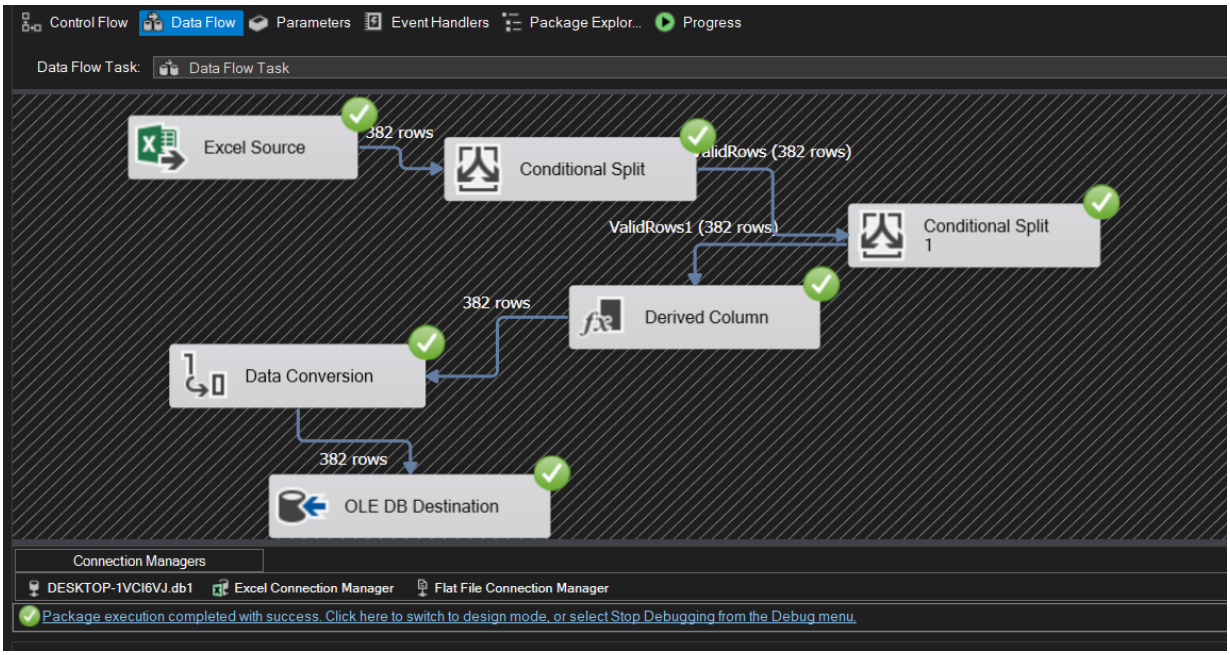
VII. Create a trigger in Azure Data Factory, to run the ETL pipeline every 20 minutes with the end date 2024/12/31.

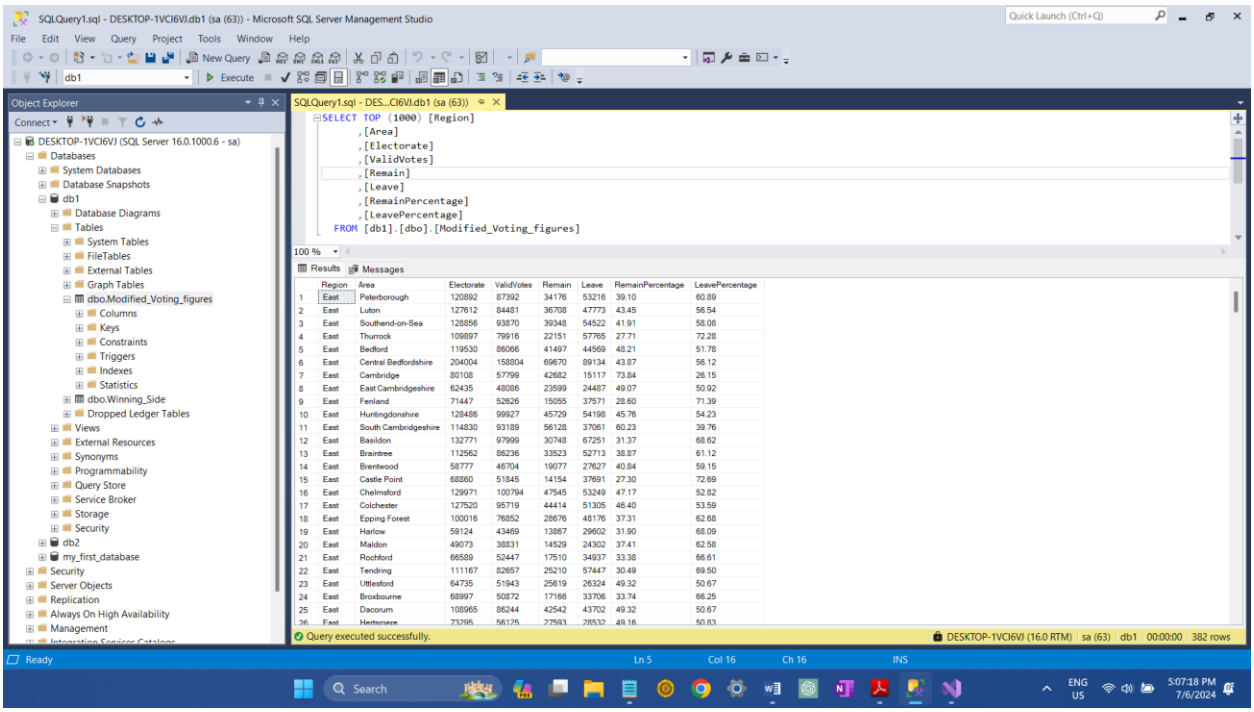

Microsoft Azure
Data Factory
geminidf1
Search factory and documentation
Home
Author
Monitor
Manage
Learning Center
Factory Resources
Pipelines
pipeline1
pipeline2
pipeline3
pipeline4
pipeline5
Change Data Capture (preview)
Datasets
Data flows
Power Query
Activities
Move and transform
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
New trigger
trigger1
Description
Type *
Schedule
Start date *
7/7/2024, 1:53:32 AM
Time zone *
Sri Jayawardenepura (UTC+5:30)
Recurrence *
Every
20
Minute(s)
Specify an end date
End On *
12/31/2024, 1:53:32 AM
Annotations
New
OK
Cancel

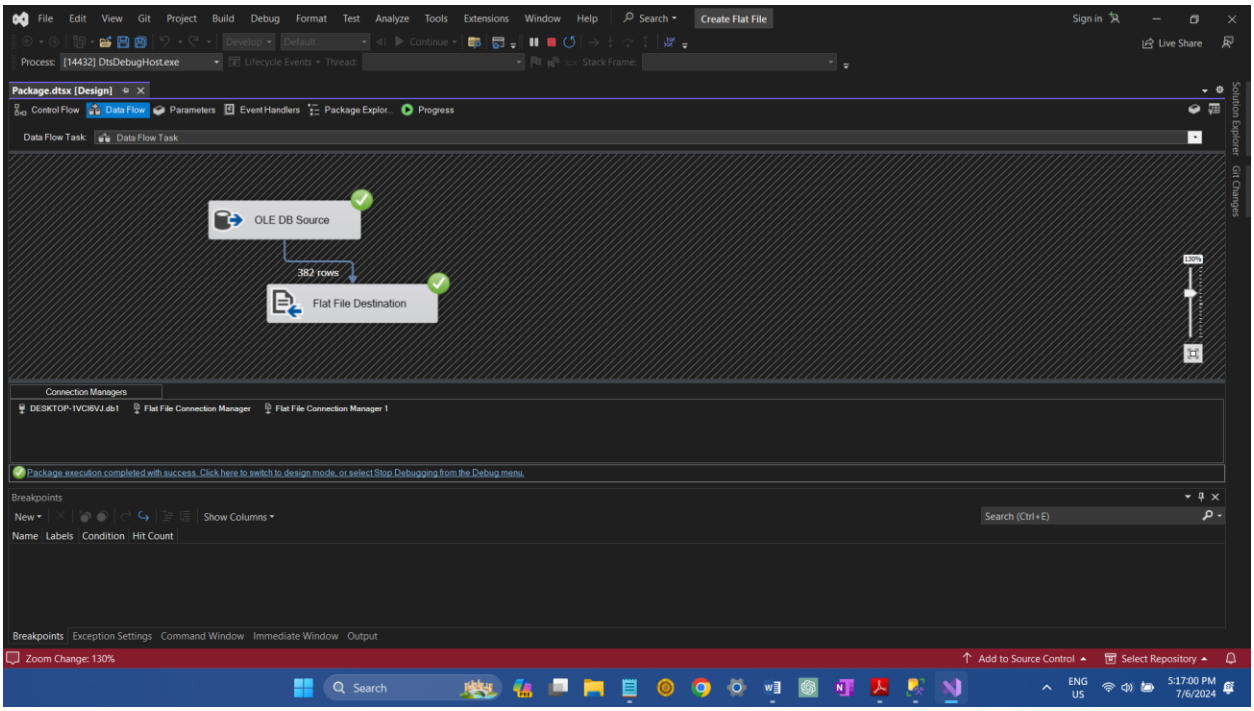

Preview experience
dataflow3
pipeline5
dataflow4
AzureSqlTable25
Validate
Debug
Trigger (1)
Data flow debug
Data flow
Data flow1

SECTION C: SSIS

Using SSIS, clean the data by removing any rows where ‘Electorate’ is less than ‘ValidVotes’ as this may indicate a data entry error. Then, calculate the percentage of votes that were for ‘Remain’ and ‘Leave’ in each area. (Hint: Get the output to a text file using the flat file destination task in the SSIS toolbox)

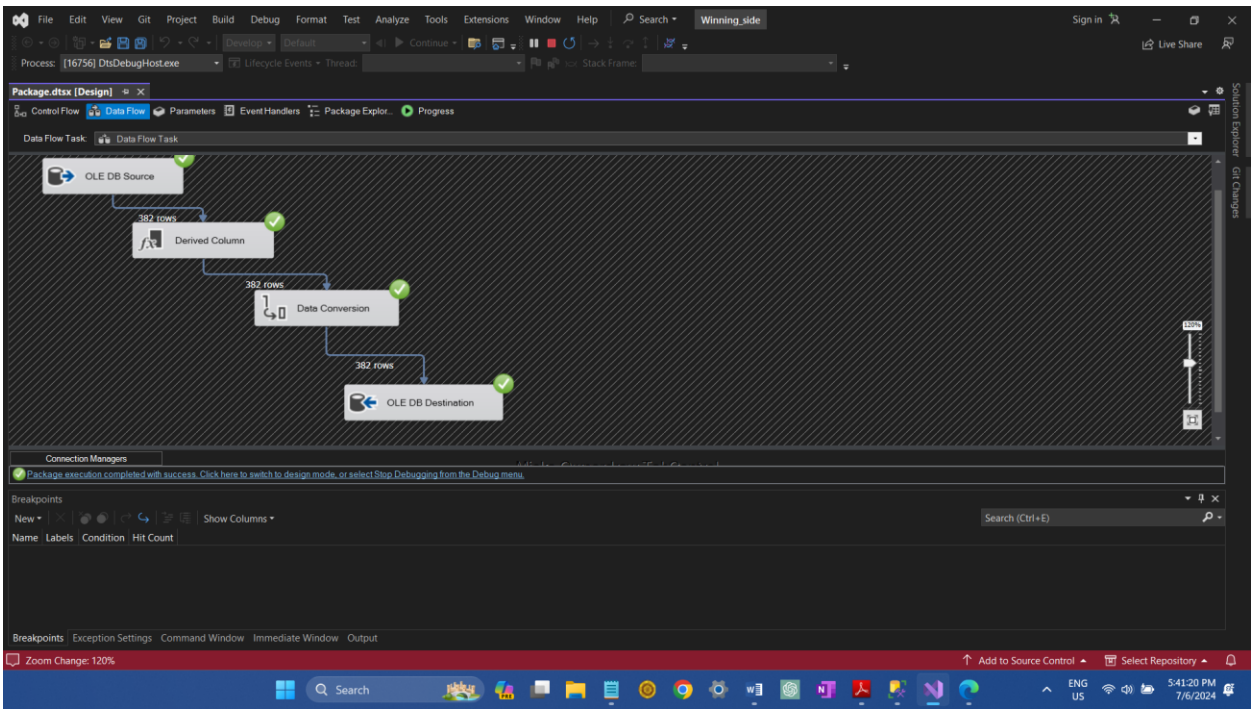


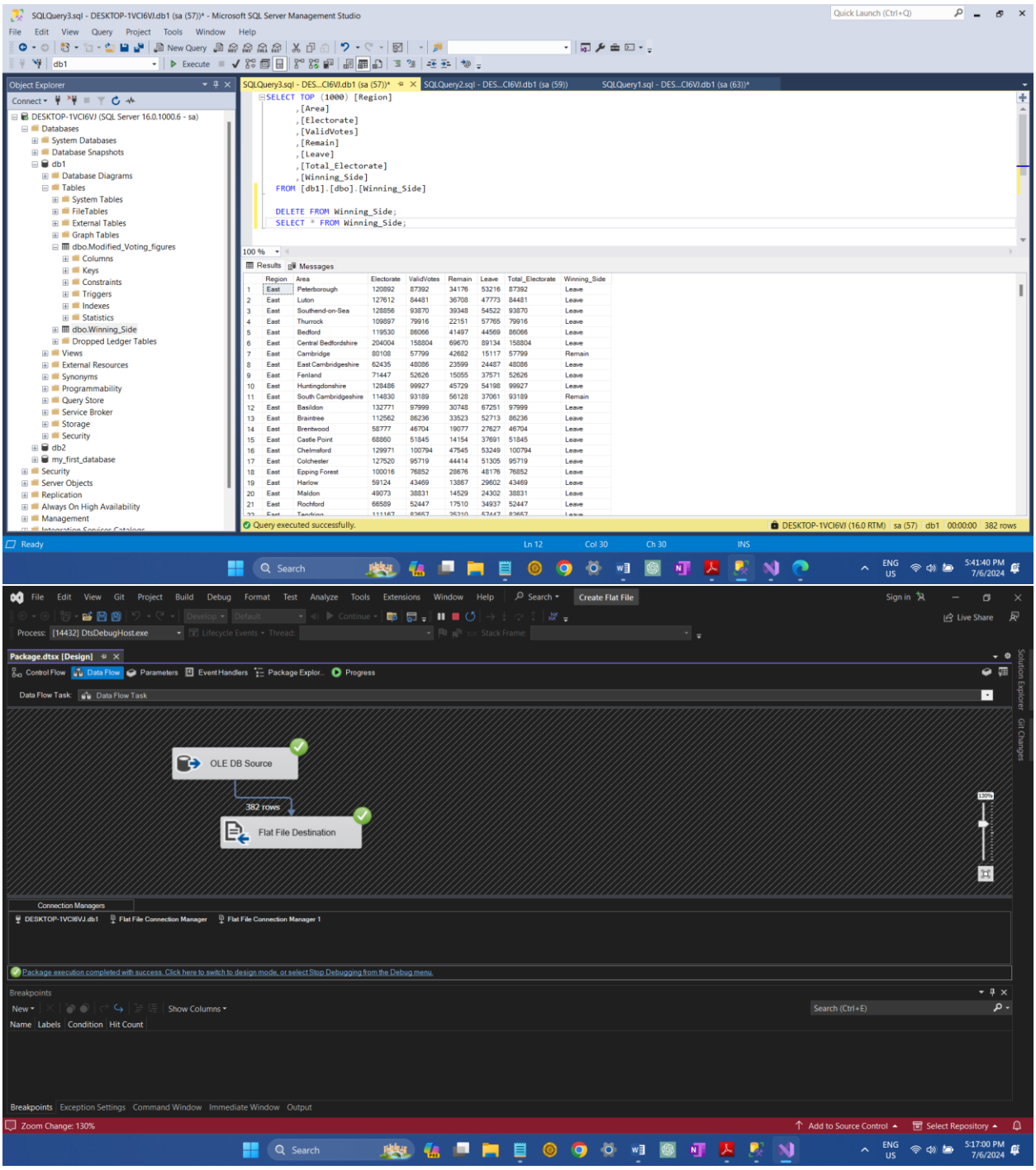




Flat File Brexit_Voters_Figure_TEXT.txt is attached

X. Create an SSIS package that aggregates the total number of votes (Remain and Leave) per region and calculates the total electorate per region. Also, determine which side won in each region based on the total votes. (Hint: Get the output to a text file using a flat file destination task in the SSIS toolbox)





Flat File Winning_Text.txt is attached.