CUSTOMER DATA MANAGEMENT AND ANALYSIS

D E P I

TEAM MEMBERS
GROUP 1

احمد علي العجمي
محمد علاء الحبال
محمد اشرف عبد العظيم
حسن محمد حسن المحلاوي
هاجر محمد ابراهيم حسن
عمر مدحت ابو شوشة

WEEK 1

SQL Queries for Data

Purpose of the Task: The goal is to perform a series of SQL queries to manage, extract, and analyze customer data.

Categories of Queries:

- Data Extraction Queries
- Data Update Queries
- Data Analysis Queries
- Bake Analysis Queries (combining multiple aspects of analysis)

WEEK 1

- **Data Extraction Queries**

SQL queries used to retrieve important customer data from the database.

a. Extract all customer information

b. Extract the top 10 customers with the highest total revenue

c. Extract the average revenue per user (ARPU) for each region

```sql
SELECT * FROM telecom_customer_churn;

SELECT TOP 10 customer_id, total_revenue
FROM telecom_customer_churn
ORDER BY total_revenue DESC;

SELECT City, AVG(total_revenue) AS ARPU
FROM telecom_customer_churn
GROUP BY City;
```

WEEK 1

- **Data Update Queries**

SQL queries for updating the existing data to reflect changes in customer status or revenue.

a. Update the churn status of a customer

b. Update the total revenue of a customer

```
UPDATE telecom_customer_churn
SET Married = '0'
WHERE customer_id = '0003-MKNFE';

UPDATE telecom_customer_churn
SET total_revenue = 1000
WHERE customer_id ='0002-ORFBO';
```

WEEK 1

- **Data Analysis Queries**

SQL queries that analyze customer behavior, revenue, and other important metrics.

a. Analyze the churn rate by region

b. Analyze the average revenue per user (ARPU) by plan type

c. Analyze the correlation between total revenue and churn status

```sql
SELECT City, COUNT(CASE WHEN Customer_Status = 'Yes' THEN 1 END) AS Churned_Customers,
       COUNT(CASE WHEN Customer_Status = 'No' THEN 1 END) AS Retained_Customers,
       COUNT(*) AS Total_Customers
FROM telecom_customer_churn
GROUP BY City;

SELECT Phone_Service, AVG(total_revenue) AS ARPU
FROM telecom_customer_churn
GROUP BY Phone_Service;

SELECT Churn_Category, AVG(total_revenue) AS Average_Revenue
FROM telecom_customer_churn
GROUP BY Churn_Category;
```

WEEK 1

- **Bake Analysis Queries**

More complex queries that combine different data points for deeper insights.

a. Analyze the top 10 customers with the highest total revenue and churn status

b. Analyze the average revenue per user (ARPU) by region and churn status

```sql
SELECT TOP 10 customer_id, total_revenue, Churn_Category
FROM telecom_customer_churn
ORDER BY total_revenue DESC;

SELECT City, Churn_Category, AVG(total_revenue) AS ARPU
FROM telecom_customer_churn
GROUP BY City, Churn_Category;
```

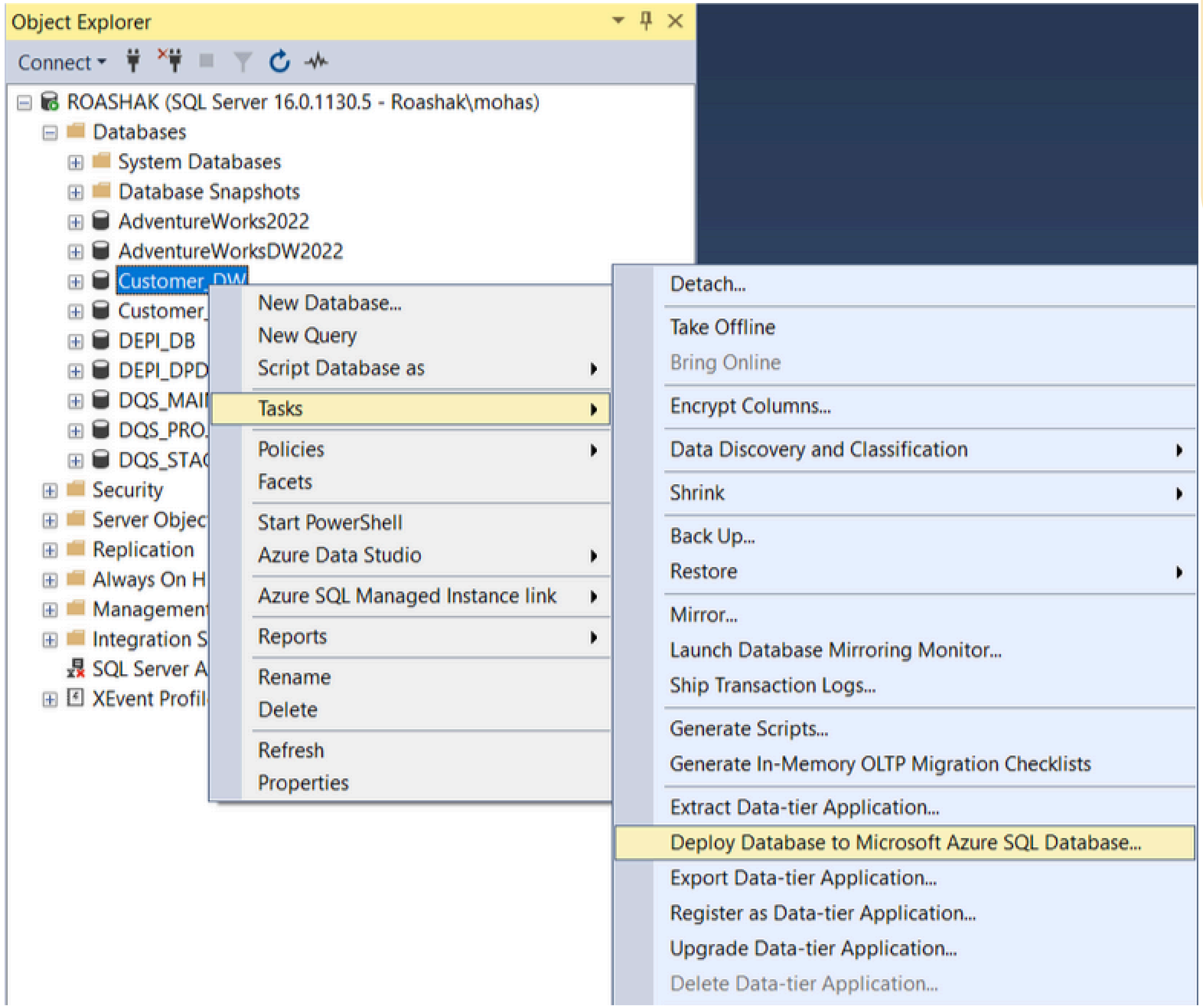
WEEK 2

- deploying the Data ware house on azure
- deploying the Data ware house on azure
- building the pipeline on the cloud

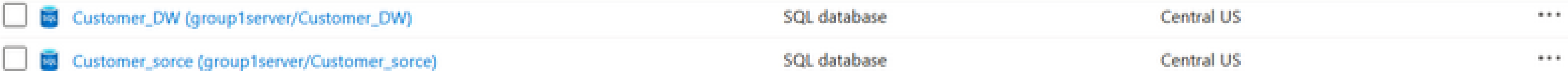
SCHEMA

dim_Subscription
subscriper_id
Customer_Status
Churn_Category
Churn_Reason
Payment_Method
Offer
dim_services
service_id
Internet_Service
Internet_Type
Online_Security
Online_Backup
Device_Protection_Plan
Premium_Tech_Support
Streaming_TV
Streaming_Movies
Streaming_Music
Unlimited_Data
Phone_Service
fact_charges
customer_id
location_id
service_id
person_id
subscriper_id
Monthly_Charge
Total_Charges
Total_Extra_Data_Charges
Total_Refunds
Total_Long_Distance_Charges
Total_Revenue
Avg_Monthly_Long_Distance_Charges
Avg_Monthly_GB_Download
dim_location_info
person_id
city
Zip_Code
Latitude
Longitude
dim_personal_info
person_ID
Gender
Age
Married
StartDate
EndDate
IsCurrent

DEPLOYING

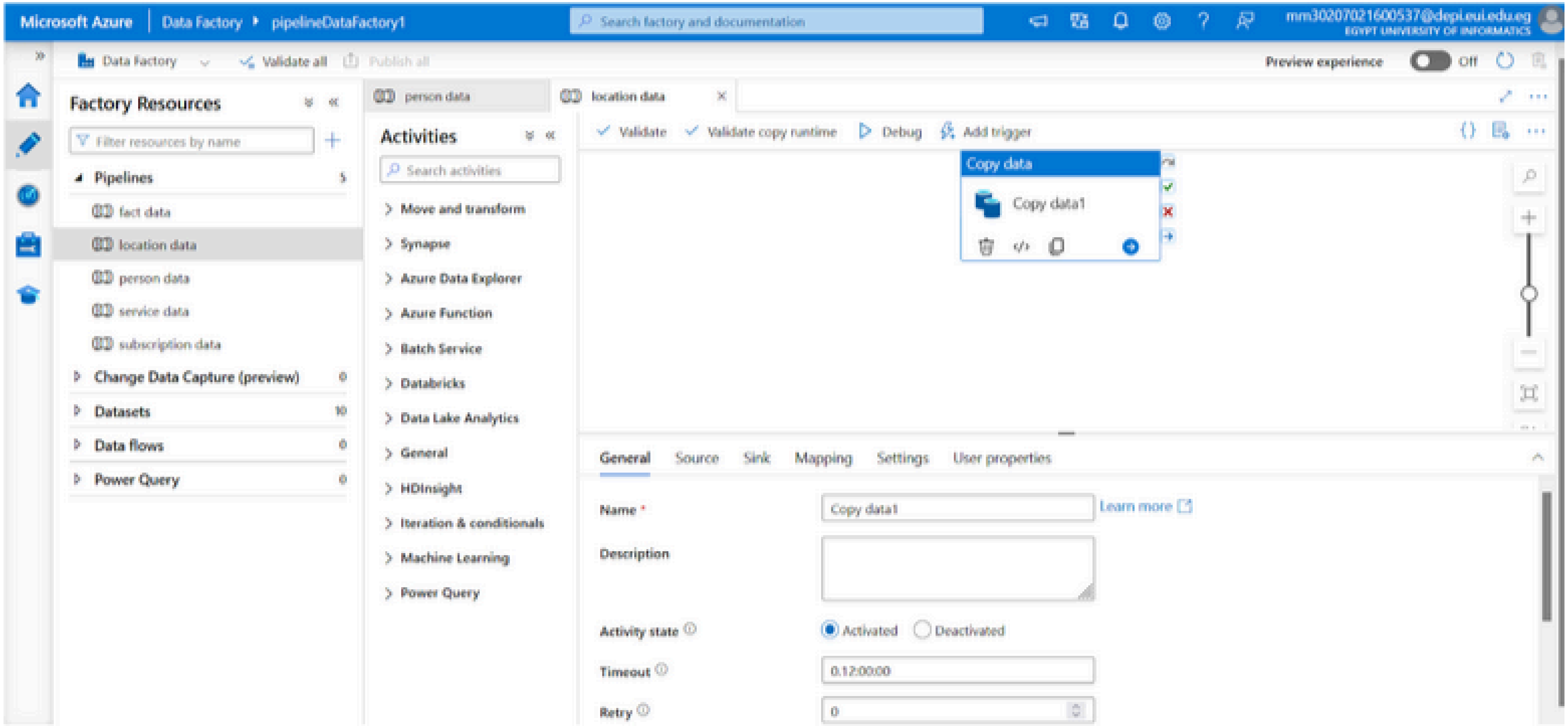


Object Explorer
Connect
ROASHAK (SQL Server 16.0.1130.5 - Roashak\mohas)
Databases
System Databases
Database Snapshots
AdventureWorks2022
AdventureWorksDW2022
Customer_DW
Customer_
DEPI_DB
DEPI_DPD
DQS_MAI
DQS_PRO
DQS_STA
Security
Server Objec
Replication
Always On H
Managemen
Integration S
SQL Server A
XEvent Profil
New Database...
New Query
Script Database as
Tasks
Policies
Facets
Start PowerShell
Azure Data Studio
Azure SQL Managed Instance link
Reports
Rename
Delete
Refresh
Properties
Detach...
Take Offline
Bring Online
Encrypt Columns...
Data Discovery and Classification
Shrink
Back Up...
Restore
Mirror...
Launch Database Mirroring Monitor...
Ship Transaction Logs...
Generate Scripts...
Generate In-Memory OLTP Migration Checklists
Extract Data-tier Application...
Deploy Database to Microsoft Azure SQL Database...
Export Data-tier Application...
Register as Data-tier Application...
Upgrade Data-tier Application...
Delete Data-tier Application...



Customer_DW (group1server/Customer_DW)
SQL database
Central US
Customer_sorce (group1server/Customer_sorce)
SQL database
Central US

AZURE



Microsoft Azure
Data Factory
pipelineDataFactory1
Search factory and documentation
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
fact data
location data
person data
service data
subscription data
Change Data Capture (preview)
Datasets
Data flows
Power Query
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
Copy data
Copy data1
General
Source
Sink
Mapping
Settings
User properties
Name
Learn more
Description
Activity state
Activated
Deactivated
Timeout
0.12:00:00
Retry
0

WEEK 3

Machine Learning

Objective: Predict customer churn using machine learning to help telecom companies retain customers by identifying those most likely to leave.

Data Source: Data extracted from the telecom_customer_churn table in Azure SQL Server.

Tools & Technologies:

- Python: Data analysis and model building
- SQL Server: Database management
- scikit-learn: Machine learning algorithms

WEEK 3

Data Exploration & Cleaning

Data Overview:

- Key columns: Age, Monthly_Charge, Total_Charges, Churn_Status
- Categorical features: Gender, Marital Status, Churn_Category

Data Cleaning:

- Handled missing values by removing null rows.
- Encoded categorical variables using one-hot encoding.

Initial Observations:

- Customers with higher Monthly Charges show higher churn rates.

WEEK 3

Data Visualization

Correlation Analysis:

- Strong positive correlation between Monthly_Charge and Total_Charges.

Churn Analysis:

- Boxplots reveal that customers paying more tend to churn.

Demographics:

- Gender and marital status distributions are balanced.

WEEK 3

Model Building & Tuning

Initial Model: Built a Random Forest Classifier and evaluated it using cross-validation.

Refined Model: Applied Gradient Boosting with GridSearchCV for hyperparameter tuning.

Best parameters:

- n_estimators: 300
- learning_rate: 0.1
- max_depth: 3

Preprocessing: Scaled numerical features for consistency across variables.

WEEK 3

Results & Model Evaluation

Performance Metrics:

- Best accuracy achieved with Gradient Boosting: 100%.
- Precision, Recall, and F1-Score for churn prediction were all strong.

Key Takeaway: The tuned Gradient Boosting model can effectively predict customer churn, providing actionable insights for customer retention strategies.

```python
from sklearn.metrics import accuracy_score, classification_report

# Predict and evaluate
y_pred_gs = grid_search.predict(X_test)
print("Gradient Boosting Classification Report:")
print(classification_report(y_test, y_pred_gs))
```

```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       477

    accuracy                           1.00       477
   macro avg       1.00      1.00      1.00       477
weighted avg       1.00      1.00      1.00       477
```

WEEK 4

Key Steps:

- Data Loading from SQL Server
- Data Preprocessing
- Model Training with Logistic Regression
- MLOps Integration (Model Logging & Registration)

WEEK 4

Data Source:
SQL Server database containing telecom customer information.
Preprocessing Steps:

- Handling missing data
- Converting customer status into binary target (Churn/Not Churn)
- One-hot encoding categorical columns

WEEK 4

Model:

Logistic Regression

Process:

- Data split: 70% training, 30% testing

WEEK 4

Tools Used:

- MLflow: Logged model accuracy and stored the trained model
- Azure ML: Registered the model for versioning and deployment

WEEK 4

Models

Model List

Register | Refresh | Delete | Archive | Deploy | Compare (preview) | Reset view | Show latest versions only | Include archived

Search | Filter | Columns

Name	☆	Version	Type	Source	Experiment	Job (Run ID)	Created on ↓
customer_churn_model		3	CUSTOM	This workspace			Oct 18, 2024 4:29 PM
customer_churn_model		2	CUSTOM	This workspace			Oct 18, 2024 4:18 PM
customer_churn_model		1	CUSTOM	This workspace			Oct 18, 2024 4:04 PM

THANK YOU