PROJECT DOCUMENTATION

Mobile Users Data Management And Analysis

OVERVIEW

Project Name	Mobile Users Data Mangament And Analysis
Project Team Members	Soad Atef Hager Ahmed
Project Dates	Start Date: Sep 28, 2024 End Date: Oct 18, 2024
Background	The rise in mobile technology has led to vast amounts of user interaction data that can be harnessed to understand behavior patterns, preferences, and trends. Our project focuses on managing and analyzing user behavior data to uncover insights for better decision-making. **Problem Statement**: With millions of mobile users worldwide, understanding user behavior is crucial for improving services, enhancing user experience, and making data-driven decisions. However, the sheer volume of data poses challenges in storage, analysis, and interpretation.
Objectives	• **Data Management/DWH**: Efficiently store and organize user behavior data for easy access and analysis. • **Data Cleaning**: Ensure high data quality by handling missing data and duplicates. • **Data Analysis**: Identify trends, usage patterns, and behavioral insights using SQL and Python. • **Model Development**: Build a predictive model to anticipate user actions. • **Visualization**: Create visual representations of key findings.
Workflow	**Step 1.1: Project Setup** • **Objective**: Begin with understanding the overall goal—managing and analyzing the "User Behavior" dataset. • **Team Formation**: Hager Ahmed and Soad Atef were assigned the task. • **Project Timeline**: October 18, 2024 - October 20, 2024.

- **Tools Used**:
  - **SQL Server**: For data storage and database management.
  - **SSIS (SQL Server Integration Services)**: To handle ETL (Extract, Transform, Load) operations.
  - **Python**: For further data analysis and visualization.

Step 2.1.1: Dataset Overview

- **Dataset Used**: *User Behavior Dataset*.
- **Data Description**:
- User ID
- Device Model
- App Usage
- Operating_System
- Device ID
- App_Usage_Time
- Screen_On_Time
- Battery_Drain
- Number_of_Apps_Installed
- Data_Usage
- Age
- Gender
- User_Behavior_Class
- OS_ID

Step 2.1.2: Understanding Data Structure

- Initial data inspection to understand the structure, missing values, and data inconsistencies.
- Tools used:
  - **Excel/CSV file reader**: To preview the dataset.
  - **Data profiling**: A quick overview using Python (pandas) to check the number of null values, data types, and distribution.

Step 2.2.1: Creating the Database

- **Tool Used**: Microsoft SQL Server.
- **Objective**: Set up a database to store and manage user behavior data.

SQL Commands:

1. **Create Database**
2. **Use the Database**

Step 2.2.2: Creating Tables

- **Tables Created By Star Schema**:

- **user_behavior**: Stores user activity data.
- **Dim_User (As Dimension Table)**
- **Dim_Device (As Dimension Table)**
- **Dim_Operating_System (As Dimension Table)**
- **Fact_User_Device_Usage (As Fact Table)**

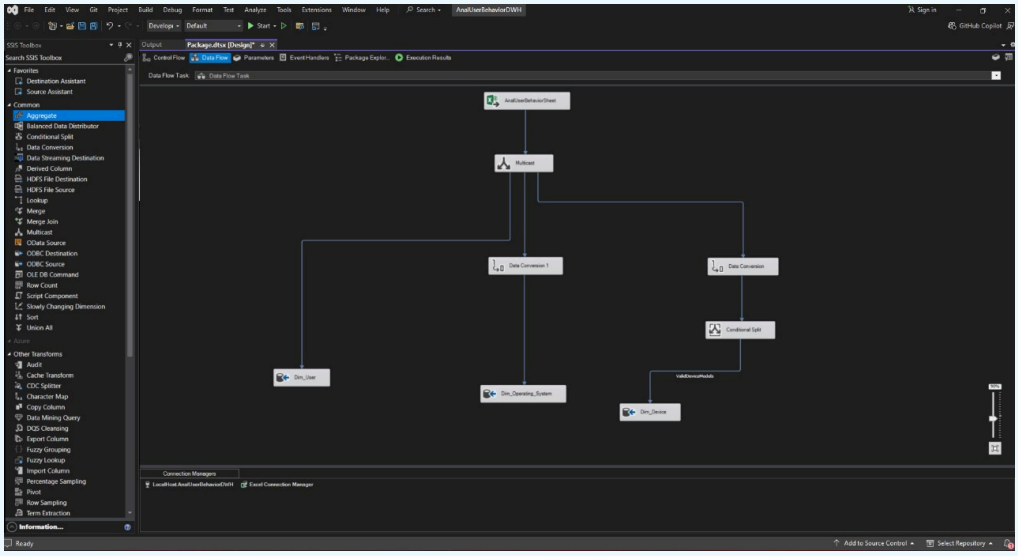
SQL Code for Table Creation

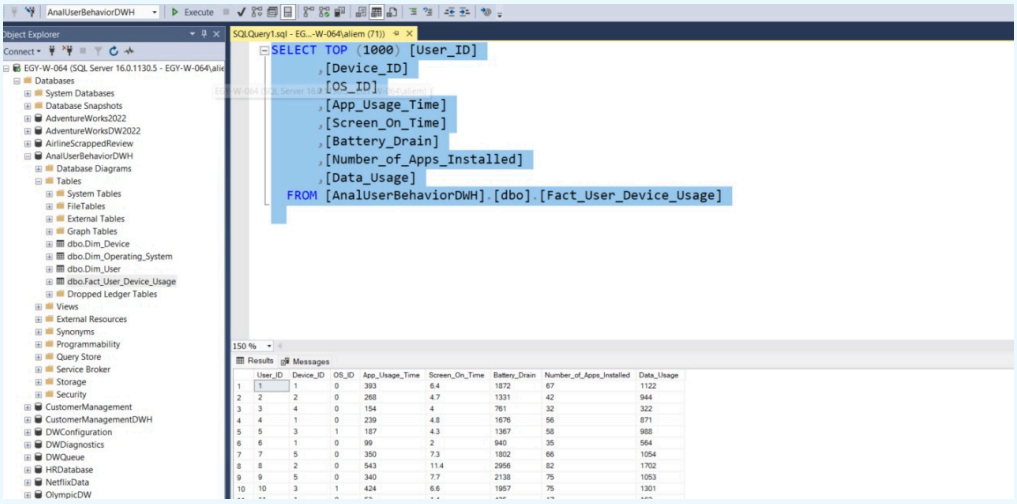
Step 2.2.3: Data Insertion

- **Objective**: Populate the user_behavior table with data from the CSV file.
- **Process**:
  - **SSIS** was used for data extraction from the CSV, transformation (data cleaning), and loading into SQL tables.

Using SSIS:

1. **Data Flow Task**: Set up in SSIS to pull data from the user_behavior.csv.
2. **Transformation**: Handle null values, clean data, and ensure all fields conform to the database schema.
3. **Load Data**: Insert clean data into the user_behavior table in the SQL database.





Step 2.3.1: Handling Missing Values

- **Objective**: Replace or remove missing data in the SQL table.

Step 2.3.2: Removing Duplicates

- **Objective**: Ensure no duplicate entries exist.

Step 2.4.1: Extracting Data from SQL Database

- **Objective**: Extract the cleaned data from the SQL server for further analysis in Python.

Python Code (using pandas and scikit learn)

Step 2.4.2: Data Analysis

- **Goal**: Analyze user behavior data to gain insights into trends and patterns.
- **Tools Used**: Python libraries like pandas and scikit learn

1. **Peak Usage Times**:
   Analyzing when users are most active
2. Device Preference Analysis

Step 2.4.3: Building a Predictive Model/ ML Algorithms

- **Goal**: Predict user engagement based on activity type

Target Audience

This project primarily benefits data analysts, marketing teams, and business decision-makers interested in understanding mobile users' behavior and product trends. The insights gained from the analysis will help optimize marketing strategies and improve user engagement.

- **Telecom Operators**: To optimize services based on user behavior.

- **App Developers**: To tailor app features and content to user preferences.
- **Marketing Teams**: To target user segments with personalized campaigns.
- **Data Analysts**: To build further predictive models on user engagement and retention.

PROJECT SPECIFICS

Project Scope	**Data Management**: Design and implement a SQL database to store mobile user data from various sources (e.g., Android and Apple products). **Data Integration**: Utilize SSIS for ETL, integrating data from multiple sources and preparing it for analysis. **Data Analysis**: Python scripting to analyze user behavior, product usage trends, and generate predictive insights.
Project Constraints	**Dataset Availability**: The analysis is confined to the 'user behavior' dataset. **Technical Limitations**: Certain advanced machine learning models may not be fully explored due to time and data constraints
Deliverables	Enumerate the specific outputs of the project: • **SQL Database**: Organized the user behavior data into an SQL database for storage and query purposes. • **Data Cleaning Report**: Document detailing how missing data and duplicates were handled. • **Data Analysis Findings**: Python scripts and reports summarizing the key insights from the dataset. • **Predictive Model**: A logistic regression model used to predict future user actions based on historical data. • **Visualizations**: Charts and graphs depicting key insights such as peak usage times, device preference, and activity distribution. • **Final Report & Presentation**: Comprehensive documentation and a presentation covering the project process, results, and recommendations.
Explorations & Decisions	**Approaches Considered** • **SQL vs. NoSQL Databases**: Initially, both SQL and NoSQL databases were considered for managing user behavior data. Ultimately, SQL was chosen due to its structured nature and the ability to run complex queries easily. • **Handling Missing Data**: Several options were considered, such as: ○ Imputation using the mean or mode.

    - Dropping rows with missing values.
    - Using advanced techniques such as KNN imputation.
      The final decision was to use simple mean and mode imputation due to time constraints and the nature of the data.
- **Model Selection**: Different predictive models were evaluated, including:
    - Logistic Regression (for binary classification).
    - Lasso
      Logistic regression was chosen for its simplicity and ease of implementation within the project timeline.
    - Ridge

Why These Decisions?

- **SQL Database**: The decision to use SQL was made because of the relational nature of the dataset, which made it easier to organize, store, and analyze the data.
- **Simple Data Cleaning Techniques**: Given the project's short timeline, we opted for straightforward data cleaning methods (e.g., mean imputation and removing duplicates) to ensure timely delivery while maintaining data quality.
- **Logistic Regression**: Chosen for its interpretability, logistic regression allowed us to quickly develop a predictive model without requiring extensive computational resources or complex hyperparameter tuning.

PROJECT TIMELINE

Task or Deliverable	*Owner*	*Notes*
Dataset Cleaning	Both	Completed successfully
Data Analysis (SQL + Python)	Both	Generated insights
Model Development & Deployment	Both	Built and deployed basic model
Visualization	Both	Final visualizations delivered

CONCLUSION

Project Outcomes

Conclusion

The project successfully managed and analyzed the user behavior data, providing valuable insights into mobile user engagement. The predictive model offers potential for telecom operators and app developers to understand user behavior and tailor services better accordingly.

Project Outcomes

- **Improved User Insights**: Detailed understanding of when and how users interact with mobile apps.
- **Predictive Model**: A basic model predicting user activity that can be further developed.
- **Actionable Recommendations**: Recommendations for marketing and service improvements based on user behavior.

Recommendations

For future work:

- **Extend the Model**: Incorporate additional datasets (e.g., demographic data) to improve the predictive model.
- **Real-time Analysis**: Implement real-time data pipelines for ongoing analysis.
- **Advanced Machine Learning**: Explore more advanced models (e.g., neural networks) to enhance prediction accuracy.