**WQD7005 DATA MINING**

2024/2025 S1

**Assignment (15%) – Individual**

You are tasked with designing an optimized data model for an e-commerce dataset using **Google Cloud AutoML** and **no-code/low-code** modern cloud-based data warehousing technologies. The dataset contains information about customers, products, orders, and order details.

**Scenario:**

You are provided with an e-commerce dataset comprising several entities:

* **Customers**: CustomerID, Name, Email, SignupDate
* **Products**: ProductID, Name, Category, Price
* **Orders**: OrderID, CustomerID, OrderDate, ShipDate
* **OrderDetails**: OrderID, ProductID, Quantity, Discount

**Requirements:**

**1. Automated Feature Engineering:**

Use **Google Cloud AutoML** to automatically generate new features from the dataset. This tool requires minimal setup and programming knowledge. Follow the step-by-step guide to upload your dataset and allow AutoML to perform feature engineering.

* Focus on how AutoML automates the generation of new features like the relationship between the product price and the discount.
* **No manual programming** is required.

**2. Star Schema Design:**

Design a simple **Star Schema** for a cloud-based data warehouse using tools like **Google Sheets** or **Airtable**. These tools allow you to visually organize your data into **fact and dimension tables** without coding.

* Use the provided **visual schema template** in a tool like **Lucidchart** to map out the relationships between tables (Customers, Products, Orders, OrderDetails).
* Instead of real-time ingestion, you can manually **upload CSV files** to Google BigQuery or Airtable to simulate the ingestion process.

**3. Data Governance and Security:**

Create a basic data governance strategy using the built-in security features of **Google Sheets** or **Google Cloud**.

* Set up access permissions for your dataset using Google Sheets' sharing options (e.g., sharing with specific people).
* Explain how this simple security model complies with data privacy regulations such as **GDPR** or **CCPA** (e.g., controlling who has access to customer data).

**4. No-Code/Low-Code ETL Pipeline:**

Instead of using complex serverless functions like **AWS Lambda**, use **Zapier** or **Integromat** to automate a basic part of your ETL pipeline. These no-code tools can move data between systems without requiring you to write code.

* Example: Use Zapier to automatically send new order data from Google Sheets to a cloud database or dashboard.
* If you want to try some programming, you can use a simple **Google Apps Script** to automate small tasks like adding new data to a spreadsheet.

**5. Data Visualization:**

Use **Google Data Studio** or **Tableau Public** to create a simple dashboard that visualizes key metrics from the e-commerce dataset. A **template** will be provided to help you get started.

* Connect your data (e.g., from Google Sheets or BigQuery) to Google Data Studio.
* Focus on creating basic visualizations like **bar charts**, **line graphs**, or **pie charts**.

**Deliverables:**

1. **Feature Engineering Report**:  
   Summarize the insights gained from using Google Cloud AutoML.

* What new features were generated automatically?
* What trends or insights were identified from the new features?

1. **Star Schema Diagram**:  
   Create a visual Star Schema using **Google Sheets**, **Airtable**, or **Lucidchart**.

* Explain how your schema organizes the dataset into **fact and dimension tables**.

1. **Data Dictionary**:  
   Provide a data dictionary with attributes for each table (e.g., CustomerID, Name for the **Customers** table). You can use **Google Sheets** or a similar tool to organize this.
2. **Data Governance Plan**:  
   Write a brief report explaining how you implemented data security using **Google Sheets’ sharing features** or **Google Cloud permissions**.

* Discuss compliance with data privacy laws (GDPR, CCPA).

1. **No-Code Automation**:  
   Explain how you used **Zapier** or **Google Apps Script** to automate a simple ETL task.

* Provide a screenshot of the automation or describe what it does.

1. **Dashboard**:  
   Submit a dashboard created using **Google Data Studio** or **Tableau Public**. The dashboard should display key insights, such as **total sales**, **top-selling products**, or **customer signups over time**.

**General Tips:**

* Use **step-by-step guides** and **templates** provided for each task.
* **Focus on concepts** (e.g., Star Schema, data governance) rather than coding. You can use built-in cloud tools to achieve the objectives.
* If you run into any technical challenges, use the **provided templates** or seek help from your instructor or peers.

**Evaluation Criteria (Total: 15%):**

1. **Feature Engineering with AutoML (3%)**:
   * Entity and EntitySet Creation (1%): Evaluate whether entities were correctly defined in AutoML.
   * Feature Synthesis (1%): Assess the quality and relevance of new features generated automatically by AutoML.
   * Insights Gained (1%): Award marks based on the student’s ability to interpret and explain the features.
2. **Star Schema Design (5%)**:
   * Fact and Dimension Tables (2%): Accurate and complete identification of tables and relationships.
   * Schema Relationships (1%): Appropriate use of relationships between tables.
   * New Features Incorporation (1%): Include relevant features generated from AutoML.
   * Scalability (1%): Consider how this schema could scale in a cloud environment.
3. **Data Governance and Security Plan (2%)**:
   * Security Strategy (1%): Clear explanation of how you controlled access to the dataset (e.g., sharing permissions).
   * Data Access Control (1%): Design role-based access using Google Sheets or other built-in tools.
4. **No-Code ETL Pipeline (2%)**:
   * Correctness of Automation (1%): Automation tasks should perform the intended function (e.g., moving data).
   * Documentation (1%): Provide a simple description or screenshot of your automation.
5. **Data Visualization and Dashboard (2%)**:
   * Dashboard Clarity (1%): Effectiveness of the dashboard in presenting key metrics.
   * Data Presentation (1%): Use of appropriate charts/graphs to communicate insights.
6. **Reflection on Modern Technologies (1%)**:
   * Use of No-Code Tools (1%): Reflect on how no-code tools (e.g., Google AutoML, Zapier) helped you achieve the assignment goals.

**Cost Management:**

* Utilize free-tier options of tools like **Google Cloud AutoML**, **Google Sheets**, **Zapier**, and **Google Data Studio**.
* Ensure you do not exceed free-tier usage limits, such as Google Cloud AutoML’s 40 hours of free training or Google BigQuery’s 1TB free query processing per month.