The Hashemite University
Prince Al-Hussein bin Abdullah II Faculty for IT
Department of Information Technology
Data Engineering and Analytics (2010042211)
Assignment 2
Building Data Pipeline and Workflow Operations
Due Date: 6-1-2025 11:59 AM
Max score: 100 points

Instructions:

1- Please keep in mind that **late submissions** will result in a **ZERO** score.
2- You must be able to discuss the details of your solution with your instructor.

Objective:

Develop a complete data pipeline using Apache Airflow to automate the ETL (Extract, Transform, Load) process, integrate machine learning, and manage workflow operations

Dataset:

The dataset to be used in this assignment is based on health and lifestyle factors that are potentially linked to diabetes. The data can be accessed via the server at the IP address 87.236.232.200 and the URL: http://87.236.232.200:5000/data .

Requirements:

Task 1: ETL Implementation

Implement the Extract, Transform, Load (ETL) process for the diabetes datasets.

1. **Data Extraction:**
   - Write a Python code to fetch the diabetes datasets from 'http://87.236.232.200:5000/data'.
   - Each request to the server provides **1,000 records** of the dataset in JSON format.
   - The dataset can be retrieved by sending a GET request to the following endpoint: http://87.236.232.200:5000/data
   - The dataset is titled "diabetes.json" and it is formatted in JSON.
2. **Data Transformation:**
   - Create a Bash script to perform at least 3 data cleaning and transforming tasks and then integrate this script into your Apache Airflow DAG using the **BashOperator (file name: DataTrans.sh).** Consider handling missing values, filtering outliers, remove duplicate records and any other necessary transformations.
3. **Data Loading:**
   - Develop a Python script to load the transform data into a MySQL database on localhost. **Please make sure that the data is integrated with the loaded data.**
4. **Machine Learning Model Integration:**
   - Implement a Python script that builds and applies a machine learning model for a classification task using the preprocessed data.

- Report the performance results of your model
- Apply the model to unseen datasets uploaded in your Mysql Local host

Task 2: Building Data Pipeline with Apache Airflow

Integrate the ETL process into an Apache Airflow DAG for scheduled execution.

1. **Apache Airflow DAG Setup:**
    - Establish an Apache Airflow DAG named 'data_pipelin'
    - Define the DAG's start date, schedule interval (set to run hourly), and other necessary configurations.
2. **Task Definitions:**
    - Define **PythonOperator** tasks within the DAG for each subtask of the ETL process:
        1. **Data Extraction**
        2. **Sensor Verification**
        3. **Data Transformation (BashOperator)**
        4. **Data Loading.**
        5. **Machine Learning Model Application**
    - Set up dependencies between tasks based on their execution order.
3. **Monitoring and Error Handling:**
    - Implement monitoring mechanisms within the Apache Airflow DAG for tracking performance.
    - Establish effective error-handling strategies to handle potential issues during execution.
4. **DAG Execution in Python:**
    - Use the Apache Airflow API or command-line interface for execution.

Submission Guidelines:

- Submit one Python script file that includes for the ETL implementation and the Apache Airflow DAG setup.
- Include comments or brief explanations within the code to enhance clarity.
- Provide any necessary instructions for setting up and running the scripts.

Evaluation Criteria:

- Correctness and functionality of both ETL and Apache Airflow scripts.
- Proper integration of tasks within the Apache Airflow DAG.
- Clarity and efficiency of the data processing, machine learning, and DAG execution.
- Effective error handling and monitoring strategies.