EM/BDO-P
Data Engineer - Assessment

Overview

As an automotive supplier, we are interested in understanding the broader trends in the automotive industry. For this task, you will create a data pipeline to acquire, process, and load data from various public sources.

Objective & Tasks

Your objective is to create a script or a series of scripts (preferably in Python or SQL) that:

1. Data Acquisition: Write scripts to download the datasets from the provided public data sources.
2. Data Processing: Clean and integrate these datasets. This should include, but not be limited to, handling missing values, duplicates, and possible outliers.
3. Data Transformation: Transform the data into a format suitable for further analysis. Justify the choices you make during this process.
4. Data Loading: Write a script to load the data into a hypothetical data storage system. While you cannot actually load the data into Azure SQL Database or Databricks Delta Lake, you should simulate the process and include the relevant commands in your script.
5. Automation Suggestion: Describe how you would automate this pipeline with a schedule interval you would choose and explain why.

Data Sources

Use the following data sources for this task:

1. U.S. Department of Transportation - National Highway Traffic Safety Administration: Vehicle Complaints
2. U.S. Department of Energy: Alternative Fuel Stations
3. U.S. Environmental Protection Agency: Vehicle Fuel Economy Information

Documentation

Provide a MS PowerPoint presentation that:

1. Explains your scripts: what each one does, and how to run them.
2. Documents any decisions you made during the data processing and transformation steps.
3. Explains your data loading process and your strategy for automation.
4. Discusses any challenges you faced and how you addressed them.

Delivery

Please upload all your scripts, the processed data, and your presentation file to a public GitHub repository and share the link with us.

Evaluation Criteria

We will evaluate your work based on:

1. Communication (30%): Clarity and completeness of your documentation.
2. Correctness (25%): Accuracy of your data processing, transformation, and loading processes.
3. Efficiency (20%): Quality and efficiency of your scripts.
4. Creativity (15%): Novel approaches to data processing and transformation.
5. Robustness (10%): Consideration of edge cases and potential errors.

Remember, while efficiency is important, it is more crucial for us to understand your thought process, your problem-solving skills, and your ability to clearly communicate your methods and results.

Good luck!