Project Overview

**Instructions**

Now that you are equipped with the knowledge and skills to extract, transform and load data you will use these skills to perform ETL, create a pipeline and upload the data into a database. You will use both Airflow and Kafka in the Hands-on labs.

**Scenario**

You are a data engineer at a data analytics consulting company. You have been assigned to a project that aims to de-congest the national highways by analyzing the road traffic data from different toll plazas. Each highway is operated by a different toll operator with different IT setup that use different file formats. In the first Hands-on lab your job is to collect data available in different formats and, consolidate it into a single file.

As a vehicle passes a toll plaza, the vehicle's data like **vehicle\_id**,**vehicle\_type**,**toll\_plaza\_id** and timestamp are streamed to Kafka. In the second Hands-on lab your job is to create a data pipe line that collects the streaming data and loads it into a database.

**Grading Criteria**

**There are a total of 30 points possible for this final project.**

Your final assignment will be graded by your peers who are also completing this assignment within the same session. Your grade will be based on the following tasks:

* Task 1.1: Define DAG arguments (2pts)
* Task 1.2: Define the DAG (2pts)
* Task 1.3: Create a task to download data (2pts)
* Task 1.4: Create a task to extract data from csv file (2pts)
* Task 1.5: Create a task to extract data from tsv file (2pts)
* Task 1.6: Create a task to extract data from fixed width file (2pts)
* Task 1.7: Create a task to consolidate data extracted from previous tasks (2pts)
* Task 1.8: Transform the data (2 pts)
* Task 1.9: Define the task pipeline (1pt)
* Task 1.10: Submit the DAG (1pt)
* Task 1.11: Unpause the DAG (1pt)
* Task 1.12: Monitor the DAG (1pt)
* Task 2.1: Start Zookeeper (1pt)
* Task 2.2: Start Kafka server (1pt)
* Task 2.3: Create a topic named **toll** (1pt)
* Task 2.4: Download the Toll Traffic Simulator (1pt)
* Task 2.5: Configure the Toll Traffic Simulator (1pt)
* Task 2.6: Run the Toll Traffic Simulator (1pt)
* Task 2.7: Configure **streaming\_data\_reader.py** (2pts)
* Task 2.8: Run **streaming\_data\_reader.py** (1pt)
* Task 2.9: Health check of the streaming data pipeline (1pt)

**How to submit**

A screenshot in JPEG or PNG format is required to be submitted for all tasks. The screenshots will be uploaded in the submission step of the final project. You will be prompted to save screenshots throughout the labs and these will be the files you submit during the Project Submission and Peer Review section of this course.