

Question1 :

Complete the class exercise and load into your snowflake table

Question2:

Use Airflow to build an ETL pipeline to use a health dataset from GitHub

(<https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv> ).

The dataset related to diabetes, which contains various health-related metrics.

- Build an ETL pipeline to store the transformed data in a snowflake Table
- .Columns names are: 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'
- Transformation: filter rows where Glucose level is greater than 120
- Clean up the temporarily downloaded file
- The pipeline should run every 1 hour

Question3:

You have been supplied with a simulated python code that produces and consumes facebook posts of different users using Kafka. Load the consumed in your destination snowflake Table

Note:

- For this project, produce a well documented report that is easy to reproduce and run. Include necessary screenshots and write up where necessary
- List out the task conducted by each member