CM 2606 Data Engineering Assignment Specification (Stage 2)	
Module Leader	Dr TMKK Jinasena
Stage	2
Unit (Group/Individual)	Individual
Weighing	40%
Qualifying Mark	40
Learning Outcomes Covered in this Assignment:	LO 3: Design and develop an appropriate Extract Transform and Load (ETL/ELT) process, using a state-of-the-art tool for a given data science requirement. LO 4: Develop and integrate a data engineering solution for a machine learning task adopting and extending methods that are informed by current research and industry practices.
Handed Out Date	17th February 2023
Due Date	7th April 2023
Expected Deliverable	Data Pipeline setup + Report + Demo/Viva
Method of Submission	Data Pipeline with respective source code (including scripts / config files) + Report in PDF format + video of the demo to be submitted online via Campus Moodle
Method of Feedback and Due Date	Rubric based, 8th May 2023
BCS Criteria (Pending) Met by this Assignment	N/A

This is an individual Assessment

- You are not authorized to work with other students or non-students, nor to share solutions for this assessment.
- This contravenes University Academic Regulations and those doing so will have been deemed to have colluded and will be subject to Academic Misconduct procedures.
- All sources must be appropriately referenced at the point of their use in the submission.

Assessment Regulations

Refer to the "How you are assessed section" in the Student Handbook for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

Coursework received late without valid reason shall not be accepted and shall receive no grade but shall count as one of the assessment opportunities prescribed in paragraph 9 of RGU Academic Regulation A4 section 4.3.

It is recognized that on occasion, illness, personal crisis or other valid circumstances can mean that you fail to submit and/or attend an assessment on time. In such cases you must inform the School of any extenuating circumstances through a **Coursework Extension Form** or a **Deferral Request Form**, with valid evidence for non-submission of an assessment up to a maximum of five working days after the assessment submission date. This information will be reported to the relevant Assessment Board that will decide whether a student should be allowed to reattempt without penalty (a deferral). For more detailed information regarding University Assessment Regulations and accessing forms, please refer to the following website: www.rgu.ac.uk/academicregulations www.rgu.ac.uk/academicregulations

Grading

Marks will be awarded for the coursework based on the provided Grading Grid. These marks will be mapped onto a grade scale from A-F as determined by the individual module coordinator.

Coursework Specification

You are required to assume the role of a data engineer and design & implement an end-to-end data pipeline. You have the freedom of picking a suitable scenario, doing your pipeline design to cater it and do the implementation on a preferred tech stack (either cloud or on-prem).

This coursework is all about how you think and apply the concepts you've learnt during the lectures into a practical use case. Therefore, you will be assessed based on how your design complements the scenario you have picked, as well as how solid your implementation is.

Final Deliverables

- Your Source Code including any scripts, config files used to set up the data pipeline. If required, add a read me file explaining what each file stands for.
- A summary report as explained in Task C below.
- Final Demo video

Task A: Choose a dataset and a suitable use case / scenario based on the data [25 Marks]

1) You may choose one of the below options to select a dataset with a clearly defined schema:
   a) Public static dataset. Few places where you can find such data are:
      i) https://archive.ics.uci.edu/ml/datasets.php
      ii) https://www.kaggle.com/datasets
      iii) https://data.world/datafiniti?tab=resources
   b) Read data from an API
   c) Web scrape data

The dataset could be a normalized one where there are multiple related entities which you may have to join using a common key depending on the usage. (e.g., Customer Orders, Customer Demographics and Product Catalog)

2) After analysing the dataset (what fields there are and any relationships between entities), think of a way how this data could be used to generate insights which would add value to a business organization. You will be awarded extra marks for the uniqueness of the scenario you pick.

You are **not required to implement this insight generation section** of the pipeline; however, your **design needs to complement the insight generation mechanism** you picked, especially when choosing storage options and their schemas, as the **output from your pipeline would be the input to insight generation section.**

You may consider (not be limited to) below options:

a) If the chosen insight generation mechanism is a BI Dashboard / Reporting, you may choose a data warehouse for storage in your design and create the required fact and dimension tables to load data from the pipeline.
b) If your chosen insight generation mechanism is a data science model generating predictions for future, you may prepare the data set to be used as the features by the data scientist and store in a suitable database / data warehouse. You will need to perform any joins / aggregations and any other transformations as required for the data science model.

Task B: Design and Implement the data pipeline [40 Marks]

1. Based on the chosen dataset and the insight generation mechanism, you need to design a pipeline which would transform and transfer data from your data source to the final sink. Your design needs to have the below components,
   - **Data Source:** This could be the data you downloaded into your local machine or stored in a database / API provisioning the data.
   - **Data Lake:** A distributed storage option to act as the staging area for the row data.
   - **Ingestion flow:** To transfer data from data source to the data lake.
   - **Data Storage option:** To act as the sink of the pipeline, from where the data supposed to be read for insight generation.
   - **ETL flow:** To transfer and transform the data from data lake to your sink. Distributed processing mechanism needs to be selected to transform / process the data.
   - **An orchestration mechanism:** Enabling your entire pipeline to run in a single execution as and when required, without running all the flows one after the other manually. The ability to schedule the pipeline execution is not a must.

2. Implement the pipeline based on your design.
   - Based on the design, all the components of the data pipeline need to be implemented using a tech stack of your choice which could be either On-Prem+ Open-Source tools or in a cloud platform with specific services that you have access to.
   - Analyze all possible tools/options you could think of and decide the best one for your solution. You are required to include this analysis / comparison in your report (explained in Task C).
   - If you hit performance issues due to the volume of data in the selected dataset or resource constraints of your machines:
     - First try to apply your knowledge in optimizations to handle the issue with your code.
     - If you still receive errors, you may opt to use a subset of the data.
     - In such case, you are required to explain the changes you made to optimize the code and the basis you used to take a subset from the original dataset in your report (Task C)
   - At least two of the following cleansing / processing steps need to be included in your ETL pipeline when transforming the data:

- Duplication Handling
- Missing Value Handling
- Corrupt data Handling
- Data Normalization
- Data Standardization
- Data Type Conversions / Formatting
- Data Aggregation

- You may add other features such as data quality checks, data validation rules, a logging mechanism, monitoring mechanism etc. to your system as appropriate to improve the completeness of the pipeline.

Task C: Provide a Summary Report [20 Marks]

Produce a summary report (3 pages maximum, excluding the cover page) on your implemented solution.

This course work is all about how you think and make decisions and this report will be your instrument to explain your thought process on selecting the dataset, selecting the insight generation scenario, design decisions and implementation details. Also make use of this opportunity to list down all the assumptions you made throughout the process.

Your report should contain the following sections:

- **Cover Page:** Containing Module code & name, title of the report, student Name, RGU ID, IIT Student ID
- **Introduction:** Summary of your end-to-end solution
- **Dataset selection:** Brief description of the selected dataset. If you selected a subset of the original data mention the basis you used for selection.
- **Insight Generation Mechanism:** Clearly explain the specific scenario you came up with at granular level. E.g., Without mentioning the mechanism as 'running a data science model' you are required to mention exactly what model you are referring to i.e., sales prediction model. Also explain how it could be achieved from your dataset.
- **Pipeline Design:** Include your **design diagram** with all components clearly named. Clearly explain all the components of your design, design decisions taken along with the justification
- **Technical Implementation:** Mention all the implementation details including the comparison of tools and frameworks, programming language etc.
- **Discussion and Conclusion:** Use this section to discuss any challenges you faced and possible enhancements.

Task D: Demo/viva [15 Marks]

You are expected to upload a compulsory 5-minute demo video of your data pipeline execution. You need to prove your identity and the originality of your work.