Assessment Brief - Coursework

Academic Year	**2024-25**
Semester	**2**
Module Number	**CM2606**
Module Title	**Data Engineering**
Assessment Method	**Coursework**
Deadline (time and date)	**18th April 2025, 12:00 p.m. (IST)**
Submission	**Assessment Dropbox in the Module Study Area in Campus Moodle. Will be open on March 17th.**
Word Limit	**explained below in deliverables**
Use of Generative Artificial Intelligence (AI) text	**you can use Gen AI but make sure you understand the concepts as you will be answering questions based on this cw in an exam.**
Module Leader	**Mohamed Ayoob**

What knowledge and/or skills will I develop by undertaking the assessment?
Building and Automating a Python ETL Pipeline with Airflow on AWS EC2
On successful completion of the assessment students will be able to achieve the following Learning Outcomes: *1. Extract data from an Open source API (Open Weather)* *2. Develop skills in building and automating ETL pipelines using Python and Airflow.* *3. Gain practical experience in data extraction, transformation, and loading processes.* *4. Develop data modelling skills.* *5. Develop configuring Security and Access control skills.*
Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Task(s) – content

Question 1:

You will have to create the following data engineering ETL task and answer questions based on this process at the examination.

1. **Setup and Configuration (20%):**
   - Create an AWS account and launch an EC2 instance/ if you have difficulties in account creation you can use your local machine.
   - Install and configure Python, Airflow, and other necessary libraries on the EC2 instance/local machine.
2. **Building the ETL Pipeline (40%):**
   - Extract current weather data from the OpenWeatherMap API using Python.
   - Transform the extracted data to the desired format.
   - Load the transformed data into an S3 bucket on AWS/ or local storage system
   - Implement the ETL process using Apache Airflow, defining tasks as operators within a DAG.
3. **Automation and Scheduling (20%):**
   - Schedule the ETL pipeline to run at regular intervals using Airflow.
   - Implement sensors in the pipeline to monitor data availability and pipeline execution status.
4. **Evaluation and Testing (20%):**
   - Evaluate the performance and accuracy of the ETL pipeline.
   - Test the pipeline with different datasets and document the results.
   - Identify any potential improvements or optimizations and suggest changes.

Question 2:

This question requires the resources provided at the end of lecture 2 (What are the design schemas of data modelling? - GeeksforGeeks) and some research. You will have to create the data model and answer questions based on this data model at the examination.

You are a data analyst working for a retail company called "SuperMart." SuperMart operates multiple stores across different regions and sells a wide range of products. The company wants to analyze its sales performance to make data-driven decisions. They have provided you with the following information:

- Stores: SuperMart has stores in multiple cities, each with a unique store ID, name, and location (city, state, and country).
- Products: The company sells various products, each with a unique product ID, name, category (e.g., electronics, clothing, groceries), and supplier.

What is expected of me in this assessment?

- Sales Transactions: Each sales transaction is recorded with a unique transaction ID, date, store ID, product ID, quantity sold, and total sales amount.
- Time: The company wants to analyze sales data by different time periods (e.g., daily, monthly, quarterly, yearly).
- Customers: SuperMart also records customer information for loyalty programs, including customer ID, name, and membership level (e.g., bronze, silver, gold).

For the above scenario.

1. Identify Dimension Tables:
   - List the dimension tables required for this scenario.
   - Define the attributes (columns) for each dimension table.
   - Explain the role of each dimension table in the data model.

2. Design the Fact Table:
   - Define the fact table and its attributes.
   - Identify the foreign keys that link the fact table to the dimension tables.
   - Explain the type of facts (e.g., additive, semi-additive, non-additive) and their significance.

3. Aggregate Tables:
   - Propose at least two aggregate tables that could be created to optimize query performance for common business questions.
   - Explain the purpose of each aggregate table and how it relates to the fact and dimension tables.

4. Business Questions:
   - Provide at least three business questions that your data model should be able to answer based on your aggregate tables.

5. Diagram:
   - Draw a star schema diagram for your data model, clearly showing the relationships between the fact table and dimension tables.

Question 3:

This question requires some of the content lectures at the latter part of the course. However you must research well in advance to complete the cw in time. Research into some of the tools and technologies provided in the specification. You can fine-tune your answers based on the lecture towards the end of the course.

HealthAnalytics Inc. is a healthcare provider migrating its patient records (including sensitive data like medical histories, insurance IDs, and biometrics) to a cloud-based data lakehouse. They aim to build an analytics platform for research but must comply with

What is expected of me in this assessment?

HIPAA and GDPR regulations. As a data engineer, design a security and access framework to protect this data while enabling authorized use. Address the following:

- Authentication vs. Authorization: Propose an authentication mechanism (e.g., MFA, OAuth) and explain how you would implement granular authorization (e.g., RBAC/ABAC) for data scientists, doctors, and external auditors.
- Data Encryption: Specify how you would encrypt data at rest (e.g., AES-256) and in transit (e.g., TLS 1.3). Justify your choices.
- Access Control: Design a policy to restrict access to sensitive fields (e.g., Social Security Numbers) using column-level security or dynamic data masking.
- Compliance & Auditing: Outline steps to ensure HIPAA/GDPR compliance (e.g., audit trails, data retention policies) and tools for real-time monitoring (e.g., AWS CloudTrail, SIEM).
- Third-Party Integration: A research partner needs API access to aggregated (non-PII) data. Describe how you would securely expose this data (e.g., API gateways, rate limiting, token-based access).

Submit a 1000-word report with concepts like architecture flowchart, RBAC matrix etc. and references to specific tools/standards (e.g., HashiCorp Vault for secrets management, Apache Ranger for policy enforcement). (The purpose of this question is to assess the hazard perception of the student when working with sensitive data)

***

Task(s) - deliverables

- *Answer all questions.*
- *Question 1 - link to a video explaining how you did your process and a document documenting your process. (max 1000 words)*
- *Question 2 - A written report explaining your data model, (1) Description of dimension tables, fact table, and aggregate tables), (2) Explanation of how the model supports the business questions amd (3) Your star schema diagram. (max 2000 words)*
- *Question 3 - Submit a 1000-word report with concepts like architecture flowchart, RBAC matrix etc. and references to specific tools/standards*

How will I be graded?

You will be answering questions based on these coursework (in addition to other theory based questions) in an examination. Be rest assured that you won't be answering specifics on the code rather you will be tested conceptually. Meaning I won't ask you to write code from scratch. You will be asked to (1) fill in the blanks of a given code, (2) asked to find faults with a given code (3) or make minor changes to a given code given a new requirement. Your exam answers and your reports you will submit will be cross-referenced for verification.

Important NOTE: There will be no deadline extensions provided to the CW in any way. The coursework can easily be done in 2 months and the submission link will be available on March 17th.

If a student:

- Makes the CW submission but fails to appear at the exam - maximum marks is 40% (the student will have to face a viva)
- Fails CW submission but appears at the exam - maximum marks obtainable is 60% (the student will have to face a viva)
- Fails CW submission and fails to appear at the exam - obviously will result in a failure
- Makes the CW submission and appears at the exam - maximum marks obtainable is 100%

The overall grade for the assessment will be calculated using the marks obtained in the exam.

A	75% - 100%
B	60% - 74%
C	45% - 59%
D	30% - 44%
E	15% - 29%
F	0% - 14%
NS	Absentees in exam.

What else is important to my assessment?

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement can be found in Appendix 2 of the RGU Assessment Policy. It provides detail on the purpose, setting and implementation of wordage limits; lists what is included and excluded from the word count; and the penalty for exceeding the word count.

What's included in the word count?

The table below lists the constituent parts which are included and excluded from the word limit of a Coursework; more detail can be found in the full Assessment Word Limit Statement. Images will not be allowed as a mechanism to circumvent the word count.

Excluded	Included
Cover or Title Page	Main Text e.g. Introduction, Literature Review, Methodology, Results, Discussion, Analysis, Conclusions, and Recommendations
Executive Summary (Reports) or Abstract	Headings and subheadings
Contents Page	In-text citations
List of Abbreviations and/or List of Acronyms	Footnotes (relating to in-text footnote numbers)
List of Tables and/or List of Figures	Quotes and quotations written within "…"
Tables – mainly numeric content	Tables – mainly text content
Figures	
Reference List and/or Bibliography	
Appendices	
Glossary	

What are the penalties?

The grade for the submission will be reduced to the next lowest grade if:

- The word count of submitted work is above the specified word limit by more than 10%.
- The submission contains an excessive use of text within Tables or Footnotes.

What else is important to my assessment?

What is plagiarism?

Plagiarism is “the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student’s work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source” (RGU 2022).

What is collusion?

“Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately“ (RGU 2022).

For further information please see Academic Integrity.

What if I’m unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation ‘A4: Assessment and Recommendations of Assessment Boards‘ sets out important information about assessment and how it is conducted across the University.