Data Management

Term Two 2023/2024

WARWICK BUSINESS SCHOOL

40% Group Assignment

2000 words

This is a strict limit not a guideline. Any piece submitted with more words than the limit will result in the excess not being marked.

AI is permitted - read more below in the T&Cs.

Overview

This extensive project is designed to simulate a real-world e-commerce data environment. You will engage in end-to-end data management, from database design to data analysis and reporting. The project involves using SQLite for database management, GitHub Actions for automation, and Quarto with R for data analysis and reporting.

Objectives

- Develop a thorough understanding of relational database design and management.
- Gain proficiency in automating data-related tasks with GitHub Actions.
- Enhance skills in data analysis and report generation using R and Quarto.

Continued…/

Part 1: Database Design and Implementation (30%)

Task 1.1: E-R Diagram Design

- **Objective**: Design a detailed Entity-Relationship (E-R) diagram for an e-commerce database. You can try to replicate the workflow of any e-commerce store that you have used and understand the workflows involved.
- **Suggested Entities**: Products, Customers, Orders, Order Details, Transactions, Suppliers, Categories, Ads.
- **Relationships**: Include one-to-many, many-to-many relationships, self-referencing relationships, and any necessary associative entities by thinking of the processes involved in an e-commerce transaction.
- **Deliverable**: A database diagram should with the Chen notation that clearly indicates primary keys, foreign keys, and cardinalities. All assumptions should be stated and provided as text. Relationships are expected to be shown using relationship sets.

Task 1.2: SQL Database Schema Creation

- **Objective**: Translate the E-R diagram into a functional SQL database schema.
- **Requirements**: Create tables with appropriate data types, constraints, primary and foreign keys. Index essential columns for performance optimization. Explain how you derived the physical schema of the database for each table that you are going to create.
- **Normalization**: Ensure the schema is normalized up to at least 3NF to eliminate data redundancy.

Continued…/

Part 2: Data Generation and Management (25%)

Task 2.1: Synthetic Data Generation

- **Objective**: Generate synthetic data that realistically simulates an e-commerce environment.
- **Requirements**: Use R to generate data. Ensure data for entities like Customers and Orders reflects realistic patterns and distributions. You can use an LLM to provide you pathways to generate data based on the description of your entity. Provide full prompt sequence.

Task 2.2: Data Import and Quality Assurance

- **Objective**: Populate your SQL database with generated data and ensure data quality.
- **Requirements**: Script the data import process. Implement checks for data quality and integrity, such as validation of email formats, checking for duplicate entries, and ensuring referential integrity.

Part 3: Data Pipeline Generation (25%)

Task 3.1: GitHub Repository and Workflow Setup

- **Objective**: Show how you used a GitHub repository to manage and version-control your project.
- **Requirements**: Include your database file, scripts, and any other necessary documents as part of the repo.

Task 3.2: GitHub Actions for Continuous Integration

- **Objective**: Automate data validation, database updates, and basic data analysis tasks using GitHub Actions.
- **Requirements**: Set up workflows that trigger on specific events like push or pull requests. Implement actions that perform tasks like running data validation scripts, updating the database with new data, and automatically running basic data analysis scripts.

Continued…/

Part 4: Data Analysis and Reporting with Quarto in R (20%)

Task 4.1: Advanced Data Analysis in R

- **Objective**: Conduct advanced data analysis on your e-commerce data.
- **Analysis Areas**: Graphs related with any quantitative insight that should be time dependent and can be updated using the data pipeline.
- **Requirements**: Use R packages like dplyr, ggplot2, and tidyr for data manipulation and visualization.

Task 4.2: Comprehensive Reporting with Quarto

- **Objective**: Create an in-depth report presenting your data analysis findings.
- **Requirements**: Your report should include data visualizations and any other statistical insights with clear narratives explaining the implications of your findings. The report should be in a format suitable for non-technical stakeholders.

Submission Guidelines

- Submit a report **no more than 2000 words** detailing your approach. Submit on the 'Group Assignment' component.
- The report should be accompanied by an **MP4 Video presentation (maximum of 10 minutes)** as a separate file where you present your pipeline as in a technical interview including difficulties and how you coped with that. Submit on the 'Group Assignment Video' component
- Your GitHub repo should be public and available for inspection.
- When required include comments in your code for clarity.

Continued…/

Marking Criteria

- **Database Design**: Accuracy and efficiency of the E-R model and SQLite schema.
- **Data Integrity and Management**: Realism and quality of generated data, effectiveness of import and normalization scripts.
- **Automation Efficiency**: Reliability and utility of GitHub Actions workflows.
- **Data Analysis Depth**: Complexity and relevance of analysis, quality of insights.
- **Reporting Excellence**: Clarity, comprehensiveness, and presentation quality of the Quarto report.

SUBMISSION DEADLINE: 12:00 (UK time) Wednesday 20th March 2024

Assignment Preparation (found in your Masters Student Handbook Section 6.2b)

Word Count Policy and Formatting (found in your Masters Student Handbook Section 6.2c)

Guidelines for Online Submission (found in your Masters Student Handbook Section 6.2e)
Your assignment cover sheet can be downloaded from this page.

ChatGPT and AI (found in your Masters Student Handbook Section 6.1e)

Please agree in advance ONE person in your group, a group representative, who will submit the final document.

Once a document has been submitted by the group representative no other members within the group will be able to submit anything for the assessment.

Your group representative must submit your completed assignment online via my.wbs. The submission deadline is visible on all group members my.wbs home page. Submission dates and times are serious deadlines which must be strictly adhered to. Your group representative must submit your document by 12:00 midday (UK time), if it is submitted later than 12:00:00 your group submission will be late and penalty marks will be incurred.

Continued…/

A completed assignment coversheet must be included as the first page of your groups script (see section 6.2b/e).

Do not include any of your group members names anywhere in the assignment, all marking is anonymous.

Please include your Group Name in the header on every page and number all pages in the centre footer, your document should be in A4 page layout, Arial font size 11pt with 1.5 line spacing and 2.54 margins (see section 6.2c).

My.wbs will only allow your group representative to upload one file.

Please submit your group work as a PDF, we will not accept PDF files of scanned documents. Your group should create the assignment in the groups chosen package (e.g., Word, PowerPoint), then convert it straight to PDF. Please ensure that your group create the PDF in advance of the deadline as technological problems will not normally be accepted as mitigating circumstances for late submission.

Please name your file as your Group Name underscore Module Code, e.g., GroupName_IB9XXX.

When submitting your group assignment your group representative will need to confirm that the work being submitted is your groups, that it has been referenced and that your group understand the University Regulations with regard to Academic Integrity.

The University recognises an increasing number of technologies such as Artificial Intelligence and that they may be applicable in your completing this assessment. The assessment brief sets out specific requirements or restrictions, and your student handbook has further guidance and advice, see section 6.1e.

Your group are reminded that the inappropriate use of such a technology may constitute a breach of University policy, such as the Proofreading Policy or Regulation 11 (Academic Integrity). If your group breach these policies, it may have significant consequences for your studies. Please make sure you all read and understand the assessment brief and how AI may or may not be used.

If a generative AI or similar is permitted and has been used your group **must** make clear why such a tool or service was used, what it was used for and the representative will be obliged to confirm that you all take joint intellectual ownership of any submitted work. As appendices, and as part of your submitted work, your group must provide screenshots of the question and the AI-generated response, alongside an explanation of how the content has been utilised, and note the relevant reference alongside each screenshot.

Continued…/

When your group submit the representative must complete (physically or electronically) a declaration. This requires the explanation of the use of any AI. Failure to disclose at the point of submission may be prejudicial in any later investigations should they arise.

If you use a generative Artificial Intelligence (AI) in the process of completing this assessment you MUST set out clearly the following:

WHY you used a generative AI

WHAT it was used for

WHICH AI was used; and

If any generated content has been used directly in this submission, if so where.

Note that this declaration does NOT contribute towards the word count for the assessment.

You will also have to confirm in your declaration that the work remains yours and you have intellectual ownership of it. You may be called for viva or other interview to demonstrate such intellectual ownership. A failure to disclose the use of AI, or the use of a misleading description of its use may have significant consequences for your studies. As a result, keeping good records of your interactions is strongly advised.

The group should double check that the correct file is being submitted. Your group are responsible for ensuring that the final version of your work is submitted.

By submitting the assignment on behalf of the group, your group representative will need to confirm and agree with the following statement:

"I declare that this work is being submitted on behalf of my group, in accordance with the University's Regulation 11 and the WBS guidelines on plagiarism and collusion. All external references and sources are clearly acknowledged and identified within the contents.

No substantial part(s) of the work submitted here has also been submitted in other assessments for accredited courses of study and if this has been done it may result in us being reported for self-plagiarism and an appropriate reduction in marks may be made when marking this piece of work."

Once your group representative declares agreement with the statement, it will not be possible to claim that your group were unaware of these requirements in the event that the work is subsequently suspected of Academic Misconduct.

After submitting, please check the correct file has successfully uploaded, by opening the document link on the submission page and scrolling through the entire document.

Continued…/

Your group representative can retract the submission before the deadline to upload an updated file, the retraction button can be found in the same place your group representative submitted your group assignment. Please upload the new document before the deadline, submitting after the deadline will attract late penalties.

End