FIT5137 Assignment 3 - S2 2023 (Weight = 30%)
Due date: Wednesday, 20 September 2023, 11:55pm

Version: 1.0 – 16/07/2023

Learning Outcomes:

LO1. To understand large data importing to Oracle from other platforms
LO2. To identify and overcome the data error during the migration process

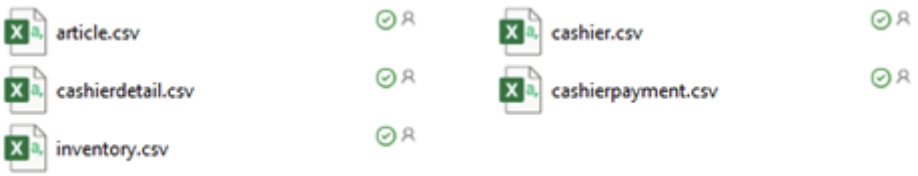
A. General Information and Submission

- This is a group assignment. One group consists of **TWO or THREE** students from the same lab you have enrolled in. You need to register your group composition through the **form** as soon as possible.
- *Submission method*: Submission is online through Moodle.
- *Penalty for late submission*: 10% deduction for each day.
- *Assignment coversheet*: You will need to sign the assignment coversheet.
- *Contribution form*: **The contribution form needs to be completed by all members and signed (e-signature is acceptable) as an agreement between members.**
- *Assignment FAQ*: There is a Assignment 3 FAQ page set up on the EdStem forum.

B. Problem Description

You have been accepted to start your internship at TelStar, an international IT company, as a Database Engineer. This company has a good reputation for handling large projects worldwide in various industries. The company has set a very high quality standard for the projects and applies this standard throughout its activities.

You have been assigned to a project called 'EPIC' where an overseas client is in the process of a complete system migration. As a member of the data transformation team, your main task is to identify the new structure for the database. The client has sent the dump file in CSV format from their previous database and the relational structure is unknown.



And it is worth noting that managers have some special requirements for the data format, as they may integrate this dataset with other data available in the company.

- The date attributes in the dataset need only include year, month and day information.
- The non-null attributes are listed in the table below. The rest of the unlisted attributes allow null values.

Table_1:Article				
articleCode	articleName	VendorKey	VendorName	TypeName
NOT NULL	NOT NULL	NOT NULL	NOT NULL	NOT NULL
Table_2:Cashier				
noTrans	dateTrans			
NOT NULL	NOT NULL			
Table_3:Inventory				
articleCode	sizes	qty	status	Barcode
NOT NULL	NOT NULL	NOT NULL	NOT NULL	NOT NULL
Table_4:Cashierdetail				
noTrans	ArticleCode	Barcode	sizes	qty
NOT NULL	NOT NULL	NOT NULL	NOT NULL	NOT NULL
Table_5:Cashierpayment				
id	noTrans	paidType	ProgressiveDisc	
NOT NULL	NOT NULL	NOT NULL	NOT NULL	

C. Tasks

Your team are required to provide the following items:

1. Create a data dictionary for each csv file. In total you will have 5 data dictionaries for the EPIC project.
    a. Output: 5 data dictionaries

Example of a data dictionary as shown below:

Table Name:	Test_Table				
Attribute Name	**Description**	**Data Type**	**Character Length/Format**	**Acceptable Null values(Y/N?)**	**Primary Key(Y/N?)**
TestID	Test_ID	Char	16	N	Y
Test_name	Name of test	Varchar2	50	Y	N

Test_score	Score of test	Number	4 digits in total 2 of which are after the decimal point	N	N
Test_date	Date of test	Date	DD/MM/YYYY	N	N
...	...	...	...	...	...
...	...	...	...	...	...

2. Create a relational diagram (ERD) that can be used to recover the data. In particular, You will need to determine the PK, FK and cardinality for each table/relationship based on the data provided and draw using the correct notation, e.g. Identifying (solid line)/Non-identifying relationships (dashed line).
    a. Output: ERD diagram

3. Elaborate on your proposed approach for data importing and data cleaning. Explain the strategies and techniques you will employ to ensure accurate and clean data.
    a. Output: Data importing and cleaning strategies.

Note: As part of the assignment, if you come across **missing or incorrect values**, it is important to document your strategy for handling them in your report. Common approaches for dealing with missing or incorrect values include:

A. Omitting rows or columns: You may choose to exclude rows or columns that contain missing or incorrect values from your analysis.
B. Imputing with mean, median, or mode: In cases where missing or incorrect values are relatively small in number, you can replace them with statistical measures such as the mean, median, or mode of the respective attribute.
C. Imputing based on neighboring values: If the missing or incorrect values exhibit some pattern or relationship with neighboring values, you can use interpolation or extrapolation techniques to estimate the missing or incorrect values based on surrounding data points.
D. Etc.

Note: Feel free to use your own approach if you think it is more appropriate.

When documenting your handling strategy, **be sure to explain the rationale for your chosen approach and its potential impact on the analysis and its limitations.** In addition, consider any domain-specific considerations or guidelines that may influence your decision-making process.

4. Maintain a detailed log of errors encountered during the data import process and demonstrate how you effectively handled each error. Clearly illustrate the errors encountered in each table and describe the steps taken to resolve them. **The aim is 100% data restoration**.
    a. Output: A SQL Script file for data cleaning
    b. Output: A report of the error, including: the error you found and how to fix it. To demonstrate the error you found, you must include the SQL queries/method you used to clean the data and the incorrect data you found.

Additional information:

Should we conduct data cleaning before importing it into the database or after the import process?

Pleas refer to Assignment 3 FAQ

5. Create a DDL script that generates the required table structures based on the defined data dictionary
    a. Output: A SQL script file for creating tables

Important Note:

- **Ensuring consistency in table names between your account and the EPIC case is essential.**
- **Prior to importing data, it is imperative that you utilize Data Definition Language (DDL) scripts to create the required tables. This step ensures that the structure of your database aligns with the EPIC scenario, facilitating a seamless data import process.**

6. Load the CSV files into the designated tables using the DDL script created above. Ensure that the data is successfully loaded into all team members' schemas for collaborative analysis.
    a. Output: No submission required for this section.6

7. Provide an SQL query to retrieve the following column information for each table:

    (1) Number of rows: The number of rows in each table.

    (2) Screenshot required.

    a. Output: A SQL Script file for retrieving
    b. Output: Screenshots of SQL query with its results

8. Conduct a descriptive analysis using SQL queries to explore the data further and present your findings in a clear and concise manner. Showcase your understanding of the dataset and highlight any noteworthy observations or patterns.
    a. Output: A SQL script file data exploration
    b. Output: Screenshots of SQL query with its results
    c. Ouput: Explanation of your findings

D. Submission Checklist

1. One **combined .pdf file** containing all tasks mentioned above:
    - □ Cover page
    - □ A signed coversheet
    - □ Details of your Oracle accounts
    - □ A contribution declaration form:

      Each student must state the parts of the assignment that they completed. An example is as follows:

      Note:
      - **The contribution percentage must end in 0, e.g, 80%, 20% [85% is not acceptable]**
      - **The example is based on a Group of 2 scenario, and the Contribution Declaration template can also be found on the Assignment 3 FAQ page, Ed forum.**

      Example:

      Percentage of contribution:

      1. Name: Adam, ID: 210008, Contribution: 60%
      2. Name: Ben, ID: 230933, Contribution: 40%

      List of parts that each student completed:
      1. Adam: list the parts that Adam did
      2. Ben: list the parts that Ben did

    - □ Task C: Report (outputs 1, 2, 3, 4, 7, 8)
    - □ Task C: Screenshot of SQL file (outputs 4, 5, 7, 8)

2. **.sql files** for the following task:
    - □ Task C (SQL commands as required by outputs 4, 5, 7 and 8)

All of the above SQL files must be runnable in Oracle.

3. Zip all the sql files from #2 above, and name the ZIP folder as DL_SQL.zip
4. **Please ensure that these tables have been successfully created with data, and that the table names in your database account are same with the EPIC case. (No submission required for this point)**

E. Submission Method

1. **Upload the PDF file from Checklist #1 and the ZIP file from Checklist #3 to Moodle by the due date:Wednesday, 20 September 2023, 11:55pm.**
   - The submission of this assignment must be in the form of **a single PDF file AND a single ZIP file**. No other forms will be accepted.
   - One member of your group can upload the submission. However, **please note that all group members must click the submit button and accept the submission statement** (failure to do so will mean your assignment will not be submitted and will incur late penalties).
   - You must ensure that you have all the files listed in this checklist before submitting your assignment to Moodle. Failure to submit a complete list of files will lead to mark penalties.

2. Penalty for late submission: 10% deduction for each day, including weekends

3. Submission cut-off time: **Wednesday, 27 September 2023, 11:55 pm**. The submission link will be unavailable after this time.

F. Late Penalty

Late assignments submitted without an approved extension may be accepted up to a maximum of **seven days** with the approval of the Chief Examiner and/or Lecturer but will be **penalised at the rate of 10% per day (including weekends and public holidays)**. Assignments submitted more than seven days after the due date will receive a zero mark for that assignment and may **not receive any feedback**.

Please note(late penalty and extension) :

1. An inability to manage your time or computing resources will not be accepted as a valid excuse. (Several assignments being due at the same time are a fact of university life.)

2. Group issues, hardware failures, whether of personal or university equipment, are not normally recognised as valid excuses. Failure to back up assignment files is also not recognised as a valid excuse.

G. Authorship

This assignment is an **group assignment** and the final submission must be identifiably your own group work. Breaches of this requirement will result in an assignment not being accepted for assessment and may result in disciplinary action.

As per the University's policy on the guidelines and practice pertaining to the usage of Generative AI, this assignment restricts all use of generative AI. In this assessment, **you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.**

H. Special Consideration

From this semester onwards, students will no longer seek extensions from the Chief Examiner/Unit Teaching Team. All extensions / special considerations will now be handled by the central Spec Con team. **Please do not email teaching staff to request an extension or special consideration.**

Extensions and other individual alterations to the assessment regime will only be considered using the University Special Consideration Policy. Students should carefully read the Special Consideration website, especially the details about what formal documentation is required.

All special consideration requests should be made using the Special Consideration Application.

Please do not assume that submission of a Special Consideration application guarantees that it will be granted – you must receive an official confirmation that it has been granted.

I. Getting help and support

What can you get help for?

- ***Consultations with the Teaching Team***

Talk to the Teaching Team: https://lms.monash.edu/course/view.php?id=162086&section=2

- ***English language skills***

Talk to English Connect: https://www.monash.edu/english-connect

- ***Study skills***

Talk to a learning skills advisor: https://www.monash.edu/library/skills/contacts

- ***Counselling***

Talk to a counsellor: https://www.monash.edu/health/counselling/appointments

J. Plagiarism and Collusion:

Monash University is committed to upholding standards and academic integrity and honesty. Please take the time to view these links.

Academic Integrity Module

Student Academic Integrity Policy

Test your knowledge, collusion (FIT No Collusion Module)

All the best for your Assignment!