Instructions on Assessment:

- You are expected to produce a word-processed answer to this assignment. Please use Arial font and a font size of 12 for text. For SQL code and output, you can use courier new font and a minimum size of 10, which preserves SQL format and layout. Where necessary, screenshots of SQL output may be used instead of plain text.
- You are required to use the Harvard Style of referencing and citation. The *"Cite them right"* guide is recommended for *referencing and citation* (Pears and Shields, 2008) which should be followed throughout your answer especially Part 5. Please do not include references to lecture notes.

Criteria for success:

For textual components :

80-100% - The description will excellently cover all the specific topics requested. The written work will be fluent, clearly presented and of out-standing quality.

70-79% - The description will comprehensively cover all the specific topics requested. The written work will be fluent and clearly presented and of distinctive quality.

60-69% - The students will show a very good knowledge of the specific topics, with very good presentation skills and quality.

50-59% - The students will show an above average knowledge of the specific topics, with above average presentation skills and quality.

40-49% - There will be an inadequate description of a significant proportion of the topics requested. There will be no major failures in presentation clarity though partly inadequate.

Less than 40% - There will be little or no information conveyed in an intelligible manner on the specific topics requested.

For SQL and other database technical components:

80-100% - The students will produce exceptional models and solutions, and will demonstrate the use of notation/language, which have outstanding syntactic accuracy (e.g., following sound algorithms, standards, methods, error free SQL code) with exceptional semantic relevance (e.g., are relevant to the requirements of the particular scenario).

70-79% - The students will produce fully complete models and solutions, and will demonstrate the use of notation/language, which have high syntactic accuracy, with high semantic relevance.

60-69% - The students will produce almost models and solutions, and will demonstrate the use of notation/language, which have appropriate syntactic accuracy with reasonably well semantic relevance.

50-59% - The students will produce fairly complete models and solutions, and will demonstrate the use of notation/language, which have adequate syntactic accuracy with reasonable semantic relevance.

40-49% - The students will produce models and solutions, and will demonstrate the use of notation/language, which have some syntactic accuracy and semantic relevance but on balance inadequate as a whole.

Less than 40% - The students will not produce sufficient models and solutions, and/or will be unable to demonstrate the use of notation/language with significant syntactic accuracy and/or significant semantic relevance.

Assessment Background and Scenario

Questions in Parts 1, 2, 3 and 5 in this assessment are based on the *TravelPortal* scenario as described in the Appendix 1. For Part 4, you should investigate the Appendix 2 for details of the relevant scenario.

Assignment Questions

Part 1 (34 marks)

(A) Using a database design approach of your choice, produce a logical design for the database to support the information system, which is needed at TravelPortal.

(24 marks)

Your answer must consist of **ONE** of the following:

- *An entity-relationship (ER) diagram (10 marks) and its mapping into a set of relations (10 marks). The ER diagram should show all relevant entity types, relationship types, attributes, keys, and structural constraints. Note that not all keys are identified/mentioned in the scenario, so you are required to identify/devise appropriate keys for all the entity types. Your ER diagram must not show any foreign keys.*

  As part of the mapping process, for each relation, you should identify appropriate primary keys as well as foreign keys (if applicable). Furthermore, you need to make sure your relations obtained from mapping your ER diagram are in 3rd normal form.

- *A set of normalised relations (10 marks) obtained through applying the normalisation process (10 marks) instead of ER modelling. You should make clear how the normalisation process has been carried out, and the reasoning employed, quoting evidence (series of steps) to support the decisions made and how various relations have been derived. Each final relation in your answer should be in 3rd normal form.*

Points to consider while preparing your solution to this part follows.

Regarding Part 1 (A):

- *The ER diagram should show all relevant entity types, relationship types, key attributes, primary keys, and structural constraints.*
- *Not all keys are identified/mentioned in the scenario, so you are required to*

identify/devise appropriate primary keys for all the entity types.

- *The ER diagram must not show/include any foreign keys or any such attributes that represent foreign keys as these are logical and not conceptual concepts.*
- *As part of the mapping process, for each relation, you should identify appropriate primary keys as well as foreign keys (if applicable). Furthermore, you need to make sure your relations obtained from mapping your ER diagram are in the 3rd normal form.*
- *You need to research, choose, and justify an appropriate naming convention for relations, attributes, and keys. You need to use your chosen convention consistently throughout your assignment. You need to document all elements of your logical relational schema in a data dictionary, in a tabular form and must be presented as text rather than an image or picture. The data dictionary should include all relevant names, descriptions, and constraints. (4 marks)*

(B) Based on your logical design from Part 1 (A) and the information available in the scenario, produce an SQL script file using Oracle 11g/12c/19c.

(10 marks)

Points to consider while preparing your solution to this part:

- An SQL script file containing appropriate SQL DDL (e.g., CREATE TABLE, ALTER TABLE, etc.) statements for creating all the relations from Part 1 (A).
- The output from running the script file in a live Oracle 11g/12*c/19c* session (e.g., using screenshots). If appropriate output is not provided, then 2 marks will be deducted from the total marks of this sub part.
- You should use relational features from the SQL92 standard in Oracle *11g/12c/19c* for constructing your data structures / tables, including appropriate primary and foreign keys.
- You should aim for a high degree of reliability in the data with the use of as many constraints as possible, e.g., check constraints on various columns (e.g., format of primary key values, positive physical values as described in the module on integrity constraints).

Part 2 (16 marks)

(A) Populate the TravelPortal database with some data (e.g., you should generate your own dummy data and load it into the TravelPortal database, consider 5 to 8 rows for each table and enough data to see meaningful output for the queries below).

(8 marks)

(B) Answer the following queries (retrievals) using Relational Algebra and SQL.

(8 marks)

q1) Display details of services for travelling between Newcastle and Penrith with 10 or more available seats in the next 14 days.

q2) Display details of the travel agent(s) with the most ticket sold in the month of March 2023.

Points to consider while preparing your solution to Part 2:

- Provide appropriate SQL DML code (e.g., INSERT) for populating the tables you have created in Part 1 (C).
- Relational Algebra expressions for Part 2 (B) (4 marks in total).
- Provide SQL queries code (e.g., SELECT) for Part 2 (B) (4 marks in total).
- Provide output for running each of above the two solutions / scripts (SQL DM Code and SQL queries) in a live Oracle 11g/12c/higher session (e.g., using SPOOL, etc.). If any of the output is missing, 2 marks will be deducted from the above. If outputs are incomplete or inadequate or misleading, then adequate marks up to a max of 4 will be deducted.

Part 3 (20 marks)

(A) Create an object-based subset of the TravelPortal database using object-relational (O-R) features of Oracle 11g/12c/19c. Select and justify any two entity types / relations of your choice which have relationships with each other, and design and implement them using nested-relational and object-relational approach covered in this module. Your answer should include object types, object tables, data loading into object tables, and answering a suitable sample query.

(12 marks)

(B) Create a NOSQL subset of the TravelPortal database using MongoDB. Select and justify any two entity types / relations of your choice which have relationships with each other and design and implement them using NOSQL database approach covered in this module. Your answer should include creating and populating collection(s) and answering a suitable sample query.

(8 marks)

Part 4 (20 Marks)

This part is based on the UNITED CREDIT CARD company's customers scenario as described in Appendix 2. The main purpose of this part is to correctly predict if credit card customers will default on their due payments. You are required to perform the following tasks:

1. Explore the dataset and justify whether UNITED CREDIT CARD company's problem belongs to predictive or descriptive data mining models. Choose which data mining task (e.g., classification, association rules, clustering, regression, etc.) will be used to produce data mining models for the UNITED CREDIT CARD company's scenario.

   (2 marks)

2. Prepare and setup your views and tables under your DMU account for accessing the shared `UnitedCreditCards` dataset, which also includes splitting the dataset for building, testing, and applying the data mining models.

   (3 marks)

3. Using the PL/SQL Data Mining API, develop at least TWO models using suitable algorithms for performing your chosen data mining task on the `UnitedCreditCards` dataset.

   (8 marks)

4. Evaluate capabilities of the models you have developed for this task.

   (3 marks)

5. Present and discuss your findings and make recommendations to the Managing Director of the UNITED CREDIT CARD company.

   (4 marks)

Note: *In order to use ODM, you should use the DMUDLn Oracle Data Mining Account allocated to you (where 1 <= n <= 65, e.g., DMUDL1, DMUDL2, etc).*

Part 5 (10 marks)

Consider the TravelPortal scenario in the Appendix. Produce a report for the Managing Director TravelPortal elaborating on professional, legal, ethical and security issues that need to be considered and make recommendations that you think are appropriate for the TravelPortal database.

(10 marks)

The report should be concise and comprehensive and in the region of 800-900 words. You should use Harvard style of citation and referencing by following the guidelines in Pears and Shields (2008).

Hand-in procedure

You should hand in your answer for this assignment as a single word-processed document electronically on the Blackboard. A sample assignment template file will be uploaded on the Blackboard near the submission deadline.

APPENDIX 1: *TravelPortal* Scenario

TravelPortal is an established transport company that has a fleet of luxury coaches/buses that serves various cities, towns, and popular tourist sites in and around England. Tickets can be bought through selected tour agents throughout the country. Tickets are obtained on first come first serve basis. Because *TravelPortal* provides excellent services for the comfort of its passengers, tickets are sold very fast.

Currently each agent has a paper-based system. There is also a paper-based system in *TravelPortal*. It is felt that it would be more efficient to have *one computer-based system*, which tracks the transactions by all agents and at the same time enable *TravelPortal* to carry out its tasks efficiently.

The current paper-based system manages logistics besides transactions in the company. Firstly, there is a sales book for the tickets sold. Each agent submits a sales list (see Figure 1) at the end of the day. The sales lists from the agents are filed in the sales book each day. Tickets are provided for each service that runs on a certain route, which has stops at certain cities/locations. Different routes may have different duration for each stop to cater for various customer needs. There is a need to record what sorts of meals are available at each stop. There may be several services for routes at different times each day (see Figure 2). *TravelPortal* constantly provides the agents with updated information on what services are available to be booked by selling tickets. The agent or its employees' phone *TravelPortal* company clerks to obtain the ticket number by specifying the service id, or if unknown, the route id, date, and time. Once the ticket is confirmed, the information is entered into the sales list.

The clerk updates the number of available seats in the service list (see Figure 2) whenever tickets are sold. The clerk also has to inform the company manager if there are no available seats for a particular service. The manager may at their discretion add an available vehicle and driver to a new service. The clerk must phone the agents to inform them if there are any updates to the list of services.

Figure 1: Sales List

Date: 03/01/2023

Agent: DaraTravel

Ticket Number	Transact Time	Service id	Price
78602152	14:36	S10243	11.50
79424931	16:04	S10565	8.90
79424987	15:05	S11125	12.90
		Total :	£33.30

Figure 2: Service List (for *TRAVELPORTAL*)

Date: 03/01/2023

Service id	Route id	Route Description	Departure Date	Departure Time	Price	No of Seats Left
S10243	NC1	Newcastle-Carlisle	04/01/2023	0900	11.50	2
S10565	DS1	Durham-Stockton	05/01/2023	1430	8.90	14
S11125	NM1	Newcastle-Middlesbrough	04/02/2023	1530	12.90	28
S38976	NP2	Newcastle-Alston-Penrith	04/02/2023	0800	25.90	50

APPENDIX 2: UNITED CREDIT CARD company's customers scenario

Dataset: UnitedCreditCards, a shared read-only table in the Oracle Database CISBG

Response variable: Default payment indicator, defaultnm

UNITED CREDIT CARD is a hypothetical credit card company operating in the UK and most of the Europe and Commonwealth countries. UNITED CREDIT CARD is having issues with many its customers (1 in 5) who are defaulting on their due payments, and causing the company to lose money, customers, and reputation. To reduce the number of customers defaulting, the company wishes to develop a model that will allow them to better predict if a customer will default within the next month, if so then to offer suitable terms to the customers that may allow them to avoid defaulting on their payments. The task is to develop few models and choose the one that can predict with highest possible accuracy/precision/recall/F1-score on whether a customer will default on their next monthly payment.

Credit Card Data Dictionary

Name	Definition
custid	Customer ID (primary key)
feature1	Credit worthiness score calculated on the basis of borrower's credit history
feature2	A score calculated based on the number and riskiness of credit enquiries made by a borrower for credit to other lenders
feature3	Severity of default by the borrower on any loan(s). Severity is a function of amount, time since default and number of defaults
feature4	Severity of default by the borrower on auto loan(s). Severity is a function of amount, time since default and number of defaults
feature5	Severity of default by the borrower on education loan(s). Severity is a function of amount, time since default and number of defaults
feature6	Minimum of credit available on all revolving credit cards (in £)
feature7	Maximum of credit available on all active credit lines (in £)
feature8	Maximum of credit available on all active revolving credit cards (in £)
feature9	Sum of available credit on credit cards that the borrower has missed 1 payment (in £)
feature10	Total amount of credit available on accepted credit lines (in £)
feature11	Amount of dues collected post default where due amount was more than 500 (in £)
feature12	Sum of amount due on active credit cards (in £)
feature13	Annual amount paid towards all credit cards during the previous year (in £)
feature14	Annual income (in £)
feature15	Estimated market value of a properety owned/used by the borrower (in £)
feature16	Number of active revolving credit cards on which full credit limit is utilized by the borrower
feature17	Number of active credit cards on which full credit limit is utilized by the borrower
feature18	Number of active credit lines on which full credit limit is utilized by the borrower
feature19	Number of active credit cards on which atleast 75% credit limit is utilized by the borrower
feature20	Number of active credit lines on which atleast 75% credit limit is utilized by the borrower
feature21	Average utilization of active revolving credit card loans (%)
feature22	Average utilization of line on all active credit lines activated in last 2 years (%)
feature23	Average utilization of line on all active credit cards activated in last 1 year (%)
feature24	Average utilization of line on credit cards on which the borrower has missed 1 payment during last 6 months (%)
feature25	Average tenure of active revolving credit cards (in days)
feature26	Tenure of oldest credit card among all active credit cards (in days)
feature27	Tenure of oldest revolving credit card among all active revolving credit cards (in days)
feature28	Number of days since last missed payment on any credit line
feature29	Tenure of oldest credit line (in days)
feature30	Maximum tenure on all auto loans (in days)
feature31	Maximum tenure on all education loans (in days)
feature32	Sum of tenures (in months) of active credit cards
feature33	Duration of stay at the current residential address (in years)
feature34	Number of active credit lines over the last 6 months on which the borrower has missed 1 payment
feature35	Number of revolving credit cards over the last 2 years on which the borrower has missed 1 payment
feature36	Number of active credit lines
feature37	Number of credit cards with an active tenure of at least 2 years
feature38	Number of credit lines activated in last 2 years
feature39	Number of credit lines on which the borrower has current delinquency
feature40	Utilization of line on active education loans (%)
feature41	Utilization of line on active auto loans (%)
feature42	Financial stress index of the borrower. This index is a function of collection trades, bankruptcies files, tax liens invoked, etc.
feature43	Number of credit lines on which the borrower has never missed a payment in last 2 yrs, yet considered as high risk loans based on market prediction of economic scenario
feature44	Ratio of maximum amount due on all active credit lines and sum of amounts due on all active credit lines
feature45	Number of mortgage loans on which the borrower has missed 2 payments
feature46	Number of auto loans on which the borrower has missed 2 payments
feature47	Type of product that the applicant applied for (C = Charge; L = Lending)
defaultnm	Indicator for default next month (1=yes, 0=no)