DUBLIN INSTITUTE OF TECHNOLOGY

DT211/4 BSc. (Honours) Degree in Computer Science (Infrastructure)

DT228/4 BSc. (Honours) Degree in Computer Science

DT282/4 BSc. (Honours) Degree in Computer Science (International)

WINTER EXAMINATIONS 2018/2019

ADVANCED DATABASES [CMPU4003]

DR. PIERPAOLO DONDIO
DR. DEIRDRE LILLIS
MS. PAULINE MARTIN – DT211C
DR. MARTIN CRANE – DT228/DT282

Thursday 10TH JANUARY   2.00 P.M. – 4.00 P.M.

TWO HOURS

INSTRUCTIONS TO CANDIDATES

ANSWER **FOUR** QUESTIONS OUT OF **FIVE**.

ALL QUESTIONS CARRY EQUAL MARKS.

Question 1 – Dimensional Model   **[25 marks]**

a. What is the *grain* in a dimensional model?   [3 marks]

b. Explain the difference between storing the full timestamp in a dimensional model versus introducing a time dimension. What are the advantages and disadvantages of the two solutions?

[4 marks]

c. The Train Ltd. company requires the designing of a data warehouse to record the number of passengers, sales on their trains. It also needs to monitor train punctuality. The starting database is composed of the following tables:

CUSTOMER (Cust_Code, Name, Address, Phone, BDay, Gender, CountyCode [FK])

TIMETABLE(Route_ID,Start_station_ID [FK], End_station_ID [FK], Dep_Time, Arr_Time, route_type)

TRAIN(Train_ID,TripDate,Dep_Time,Arr_Time,RouteID [FK])

TICKETS(Ticked_ID,Train_ID [FK], Cust_Code [FK], Num_tickets_full, Num_tickets_reduced,Date, Start_Station_ID [FK], End_Station_ID [FK], TicketClass [FK])

TICKET_PRICE(TicketClass,Date,Full_Price,Reduced_Price)

STATIONS(Station_ID,Station_Name,Latitude,Longitude,City, CountyCode [FK])

COUNTY(CountyCode,Population,Province)

The DB contains information about customers and trains. There is a timetable containing information about each train connection offered (for example the route_id=5 is "Dublin-Cork leaving at 8:40 arriving at 11:20"). The field route_type is either "daily", "weekdays" or "weekends".
The table train stores information about a single actual train trip, when the train left and arrived to the final destination and the route of the train (for instance, train with route id 5 (=Dublin Cork 8:40-11:20), left at 8:50 and arrived at 11:40 on the 6th June 2015.
The DB contains information about stations and their geographical location. Finally, the DB contains information about tickets sold for every train. Customer buys tickets by specifying the number of "full price" tickets and the number of "reduced price" tickets, and the departure and destination station. The price of each ticket is based on the date of the purchase and on the class of the ticket. The class of each ticket is computed by the system based on the departure and destination station (for instance, a Dublin-Waterford ticket is class "E"). The prices of tickets depend on the date the ticket is issued, the class of tickets and if it is a reduced or full price ticket.

Produce one or more star schema for the above ER diagram. The diagram(s) should support the following queries and reports:
(i)   A weekly report showing the total revenue for each train route
(ii)   A weekly report showing the distribution of customers by county and train route

(iii)　Show the list of trains with a number of passengers above 500

(iv)　Show the average number of tickets sold for each train in each county

(v)　Understand the demography of the customer base (gender / age)

(vi)　Show the average delay of each train route

(vii)　Show the percentage of trains that arrived at the final destination punctually(=arrival time <= expected arrival time as stored in the Timetable Table).

(viii)　Show the name of the bestselling route each week

(ix)　Show, for every month, the number of trains that arrived more than 1 hour late during the weekdays in all the counties with more than 100000 persons.

Justify all your design choices. If a field in the fact or dimension table is not in the ER diagram, explain how to derive it, where it should be derived and why.

[13 marks]

d. Using your dimensional model, write the SQL query at (viii).

[5 marks]

Question 2   **[25 marks in total]**

Consider the following information about invoices issued by a veterinary clinic to customers. The information is stored in a not normalized table:

Field	Comment
Invoice #	Unique Invoice Number
Customer Number	Unique Customer Number
Customer Name	Customer name
Customer Surname	Customer surname
Customer Address	Customer Address
Customer EIRCODE	A postal code identifying a single house
Pet ID	Unique ID of the pet
Pet Name	Pet Name
Pet Type	Dog, Cat, ….
Procedure ID	Unique ID of the medical procedure
Procedure Type	Description of the Procedure
Procedure Date	Date of the procedure
Procedure Amount	Price of the procedure
Tax	Tax to be added on top of the invoice total

Note that the same pet cannot undergo the same procedure the same day. The price of a procedure is fixed, the tax might change for each invoice. An invoice can have multiple procedures in it from different dates. An invoice can also contains procedures for different pets (but always the same owner). A pet has only one owner for its entire life.

a. How would you store the above information in a fully normalized relational diagram?

[7 marks]

b. How would you store the above information in a document-based NOSQL database (use a JSON representation)

[7 marks]

c. How would you store the above information in a Graph Database (show the nodes, links and properties).

[7 marks]

d. In general, comment *on* the advantages and disadvantages of ER model versus a JSON-like model. Provide examples to support your discussion.

[4 marks]

Question 3   **[25 marks in total]**

You are required to design a database to store information about flight connections. Each flight is described by an *flight_id*, a departure and destination airport, departure and arrival time, and airlines. Each airport is described by an address, name, international 3 digits code and country. There is a relationship among two airports *a* and *b* if there is a flight between airport a and b. Note how a single flight defines a directed link. If there is at least one flight from airport *a* to *b* we say that *a* is connected to *b*.

You are required to:

a. Provide a relational schema to store airports and flights information. Provide tables, fields and show primary and foreign keys.

[3 marks]

b. Write an SQL query to get the names of the airports that can be directly reached from the JFK airport.

[2 marks]

c. Write an SQL query to get the names of the airports that can be reached from JFK directly or within 1 stopover.

[4 marks]

d. Write an SQL query to get the names of the airports that can be reached from JFK directly or within 2 stopovers.

[4 marks]

e. Provide a JSON structure to store the same information provided in the relational model.

[4 marks]

f. Show how the same information would be stored in a graph database

[4 marks]

g. Compare the three data models: which one is easier to query? Is standard SQL a sustainable way to query the airports/flights database? Which one is the best solution for this specific problem?
Justify your answers

[4 marks]

Question 4   [25 marks in total]

a. What could be the problem with implementing indexes using binary trees? Will a write-mostly or a read-mostly application be affected from this problem? Justify your answer and provide examples

[4 marks]

b. Which are the 3 types of strategies used for managing changes in dimensions? Provide an example for each of them

[9 marks]

c. Insert in a (2,3) b-tree the following index values:

[12 marks]

50, 60, 70, 40, 30, 20, 10, 80, 90 , 100, 55

Question 5   [25 marks in total]

a.
Describe the ACID properties of a relational database.   [3]
Describe the BASE properties of a NOSQL database.   [3]
Discuss the difference between the two approaches and for which application the BASE approach is more suitable   [3]

[9 marks in total, 3 for each question]

b. Describe the concept of replica set and sharding in MongoDB. What is the main difference between these two concepts? In particular, describe the concept of primary and secondary copy, arbiter, failover, sharding key

[10 marks]

c. Describe the following configurations of a secondary copy in MongoDB:
   i. Hidden Copy
   ii. Priority 0 Copy
   iii. Delayed Copy

[6 marks]

COLLEGE EXAMINATIONS

AMENDMENTS TO EXAMINATION QUESTION PAPER

COURSE REF   VENUE:

SUBJECT: ADV. DATABASES 4003

DATE: 10/1/19

TIME:

SIGNED:

INSTRUCTIONS:

YOU CAN USE THE INTERNATIONAL 3 DIGITS CODE AS PRIMARY KEY