DUBLIN INSTITUTE OF TECHNOLOGY

DT211C BSc. (Honours) Degree in Computer Science Infrastructure

Year 4

WINTER EXAMINATIONS 2016-2017

ADVANCED DATABASES [CMPU4003]

DR. PIERPAOLO DONDIO
DR. DEIRDRE LILLIS
MR. ALAN FAHEY

TUESDAY 17TH JANUARY 1.00 P.M. – 3.00 P.M.

DURATION
2 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. An online company requires the designing of a data warehouse to record the sales of their sport sticker albums. The database is composed of the following tables:

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

COLLECTIONS (Album_Code, Collection_name, launch_date, num_pages, num_stickers, num_foil_stickers, competition_code,publisher_code [FK])

COUNTRY(Country_Code,Country_Name)

COMPETITIONS(competition_Code, comp_Name, starting_date, ending_date,Country_code [FK])

ORDER (Cust_Code [FK], Album_Code [FK], Date [FK], number_packet, number_stickers,store_code [FK])

PUBLISHER(publisher_Code, P_name,P_address,Country_Code [FK])

STORE(store_code, store_address,Country_code [FK])

PRICES(Album_code, date, packet_price,sticker_price)

Each sticker album is related to a sport competition (such as Premier League, Serie A, FIFA World Cup) and the database store the total number of pages, the number of simple and foil stickers and the publisher.
A customer can insert an order online or buy in one of the stores. The online store is just a store like the other, with a special store_code equal to 01 and using the location of the company headquarter. Customers can buy individual cards or packets, each of them containing 6 stickers. The price of stickers and packets changes – potentially every day (that is why the date is also part of the primary key of the table PRICES). Note how in the table order Album_code alone is a foreign key referencing table Collections, and album_code + date is also a foreign key referencing table Prices.

Produce a star schema for the above ER diagram. The diagram should support the following queries and reports:

(i) A weekly report showing the total revenue for each collection album or for each competition

(ii) A weekly report showing the distribution of customers by country and store name

(iii) Show the list of Publisher that sold more than "X" number of packets in each quarter

(iv) Show the average number of stickers sold for each album in each country

(v) Understand the demography of the customer base (gender / age)
(vi) Show the list of customer that did not buy any sticker in each quarter
(vii) Show the name of the bestselling album each week
(viii) Show, for every month, the percentage of the total revenue generated online versus the revenue generated in the stores

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]**

You are required to design a database to store information about employees and their managerial relationships. Each employee is described by an *employee_id*, an *employee_name* and a *salary*. There is a relationship among the employees: employee *a* is linked to employee *b* if *a* is *b*'s manager. The date employee *a* became *b*'s manager is also stored. You are required to:

a. Provide a relational schema to store employees' information and the *manager* relationship. Provide tables, fields and show primary and foreign keys.

[4 marks]

b. Write an SQL query to get the names of the employees directly managed by the employee with id=3

[4 marks]

c. Write an SQL query to get the names of the employees managed by employee with id=3 directly or managed by people managed by employee with id=3.

[5 marks]

d. Provide a *json* structure to store the same information provided in the relational model.

[4 marks]

e. Compare the two data models: which one is easier to query? Is standard SQL a sustainable way to query the employee database? What could be a better solution?

[4 marks]

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

[4 marks]

Question 3 [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

[5 marks]

b. Explain what an hash-function is. Describe one strategy for managing collision in a hashed index. What is the main problem if the hash function has not been chosen properly?

[6 marks]

c. Suppose you need to define an hash function for storing information about Irish citizens, and suppose the unique key value is the height of each person. Can you provide an example of hash function suitable for this data distribution? Justify your choice.

[4 marks]

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

[10 marks]

2, 4, 5, 6, 8, 10, 9, 14, 16, 18

Question 4 [25 marks in total]

a. What is a bitmap index and why it is used? How big (in bytes) is a bitmap index to store the field *day of the week* of a table with 5 million records?

[5 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. Normalisation and denormalisation are important techniques in the database design process. Describe each of these techniques, discussing when each might be applied.

[8 marks]

d. Provide a design situation in which a trigger is needed to guarantee the correctness of the DB implementation

[3 marks]

Question 5 [25 marks in total]

a. Explain the CAP theorem for distributed databases, providing examples of one system that drop Consistency, a system that drop Availability and a system that drop Partition-tolerance

[9 marks]

b.

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]

c. Which are the advantages of a dimensional model over an ER diagram? Which are the strengths of an **ER** Diagram?

[7 marks]