THE UNIVERSITY OF MELBOURNE

School of Computing and Information Systems

INFO20003 Database Systems

SAMPLE EXAM

Reading Time: 15 minutes

Writing time: 120 minutes

This paper has 14 pages including this page

Authorised Materials
Calculators: Casio fx82 calculators are permitted

Instructions to Invigilators

- The examination paper IS TO REMAIN in the examination room
- Students are to be provided with 1 script book
- Provide extra script books on request

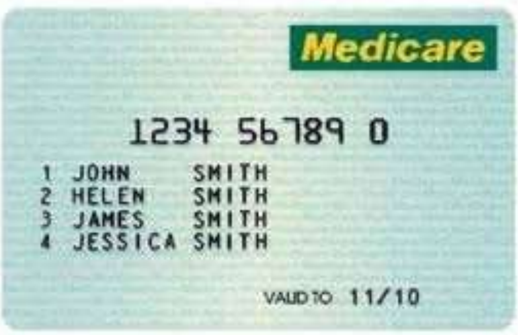
Instructions to Students

- The total mark for this paper is 120
- Ensure your student number is written on all script books and answer sheets during writing time
- Attempt all questions in all **10** sections, which are of unequal marks value
- Answer all questions on the right-hand lined pages of the script book
- Start the answer to each question on a new page in the script book
- The left-hand unlined pages of script books are for draft working and notes and will not be marked
- State clearly and justify any assumptions made
- Write legibly in blue or black pen
- Mobile phones, tablets, laptops, and other electronic devices, wallets and purses must be placed beneath your desk
- All electronic devices (including mobile phones and phone alarms) must be switched off and remain under your desk until you leave the examination venue. No items may be taken to the toilet

This page is left intentionally blank.

Section 1 – ER Modelling

The Australian Department of Health and Human Services (DHHS) has a requirement to store the vaccination record of every individual (known as a patient) that resides in Australia who is eligible for a Medicare card. For families, children under the age of 15 are stored on the family card (refer Figure 1). Each Medicare card number is associated with only one family. Every Medicare card has a 'valid to' date stored as a month and year on the card (refer Figure 1). Each family member holds a position number on the card. For example to identify Jessica Smith (refer Figure 1) both her Medicare number (1234567890) and position (4) on the Medicare card would be required.



Medicare

1234 56789 0

1 JOHN SMITH
2 HELEN SMITH
3 JAMES SMITH
4 JESSICA SMITH

VALID TO 11/10

Figure 1: Australian Medicare card (DHHS, Australia)

Each Medicare card is attached to one residential address, contact email, and phone number. For all patients listed on a Medicare card we record their gender, birthdate, first name, last name, and if they have any known allergies (e.g. Penicillin, Cortisone, Codeine). If patients do have an allergy we need to know what the allergy is and the reaction (severe, moderate or mild).

The DHHS needs to record mandatory vaccinations (e.g. Measles, Polio, Whooping Cough, Diphtheria, Tuberculosis and Tetanus) as well as optional vaccination types (e.g. HPV, Flu, Hepatitis A, Hepatitis B, Cholera, Typhoid, Yellow Fever).

For each vaccination event that is given to patients we must record the vaccination type, date of vaccination and the vaccination batch number. Vaccine producers can produce many different types of vaccines and each vaccine can have many batches.

Patients can receive their vaccination from any registered doctor. Every doctor is identified by a unique medical practitioner number (MPN). We record the Medical Practitioner's title (Dr, Mr, Mrs, Ms, Prof.), first name, last name, registered business address, email, and business phone numbers.

Q1. Draw a **conceptual model** in either Crow's Foot **or** Chen's notation for this case study (in your script book). Be sure to write down any assumptions you make.

(20 marks)

This page is left intentionally blank

Section 2 – SQL-DDL

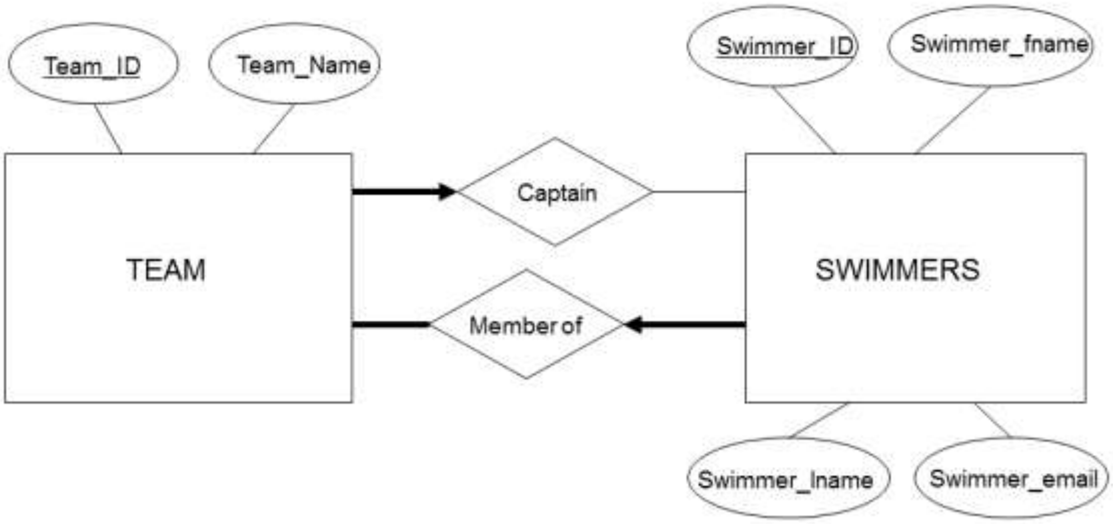


Figure 2: Data model for SQL DDL

Q2. Write SQL statements to *create* the tables for the above data model. Be sure to specify primary and foreign keys. You *do not* need to specify whether the fields are NULL/NOT NULL. Choose appropriate data types for attributes that are not obvious.

(5 Marks)

Section 3 – SQL-DML

Figure 3 shows the schema for a small business database that contains data about employees, products, customers and orders.

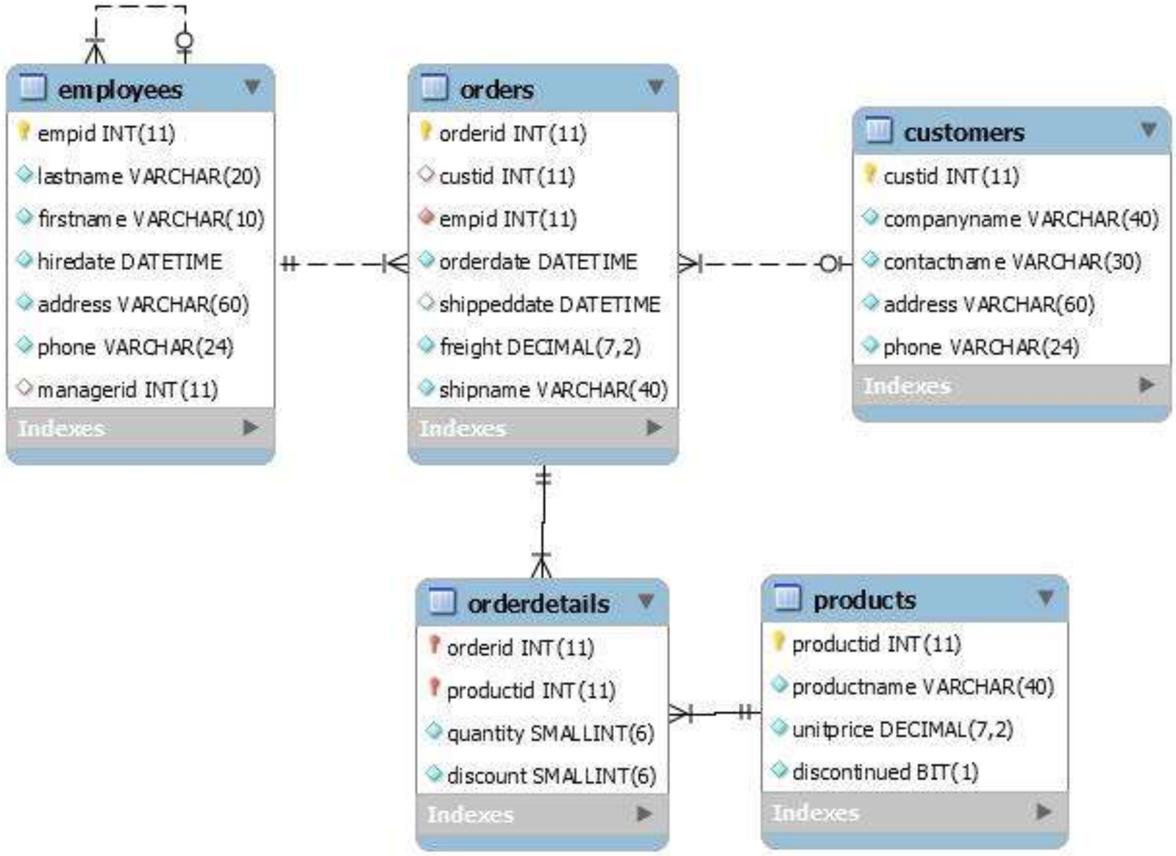


Figure 3: Data model for SQL DML

Write a single SQL statement to correctly answer each of the following questions (3A – 3D). DO NOT USE VIEWS or VARIABLES to answer questions. Query nesting is allowed.

The relations are repeated here for your convenience.

employees ($\underline{\text{empid}}$, lastname, firstname, hiredate, address, phone, $\overset{\text{FK}}{\text{managerid}}$)

orders ($\underline{\text{orderid}}$, $\overset{\text{FK}}{\text{custid}}$, $\overset{\text{FK}}{\text{empid}}$, orderdate, shippeddate, freight, shipname)

customers ($\underline{\text{custid}}$, companyname, contactname, address, phone)

orderdetails ($\overset{\text{FK}}{\underline{\text{orderid}}}$, $\overset{\text{FK}}{\underline{\text{productid}}}$, quantity, discount)

products ($\underline{\text{productid}}$, productname, unitprice, discontinued)

Q3A. Write a query that returns customers (company names) and the details of their orders (orderid and orderdate), including customers who placed no orders.

(3 marks)

Q3B. Write a query that returns the first name and last name of employees whose manager was hired prior to 01/01/2002.

(4 marks)

Q3C. Write a query that returns customers whose company name is ‘Google’, and for each customer return the total number of orders and total quantities for all products that were not discontinued (‘1’ means discontinued, ‘0’ not discontinued).

(5 marks)

Q3D. Write a query that returns the ID and company name of customers who placed orders in 2007 but not in 2008.

(8 marks)

Section 4 – Query Processing – Joins

Given the schema of Question 3 (SQL-DML), consider the relations Orders and Employees. Imagine that relation Employees has 1,000 pages and relation Orders 50,000 pages. Consider the following SQL statement:

```
SELECT *
FROM Employees NATURAL JOIN Orders
WHERE freight > 1000;
```

There are 502 buffer pages available in memory. Both relations are stored as simple heap files. Neither relation has any indexes built on it.

Q4A. What is the cost (in disk I/O's) of performing this join using the **Block-oriented Nested Loops Join** algorithm? Provide the formulae you use to calculate your cost estimate.

(3 marks)

Q4B. What is the cost (in disk I/O's) of performing this join using the **Hash Join** algorithm? Provide the formulae you use to calculate your cost estimate.

(3 marks)

Q4C. In comparing the cost of different algorithms, we count I/O (page accesses) and ignore all other costs. What is the reason behind this approach?

(2 marks)

Q4D. Which approach should be the least expensive for the given buffer size of 502 pages:

1. Simple Nested Loops Join
2. Page-oriented Nested Loops Join
3. Block-oriented Nested Loops Join
4. Hash Join

Please write the number of the correct response. No need to provide formulae for question 4D.

(2 marks)

Section 5 – Query Processing – Indexing

Given the schema of Question 3 (SQL-DML), consider the relation OrderDetails. Imagine that the relation OrderDetails consist of 100,000 tuples stored in pages and each page stores 100 tuples. Imagine that the *quantity* attribute can take any value between 0 and 20 ([0, 20]), and imagine that *discount* can take any value between 0 and 100 ([0, 100]). Suppose that the following SQL query is executed frequently using this relation:

```
SELECT *

FROM OrderDetails

WHERE quantity > 15 AND discount > 90;
```

Your job is to analyse the following query plans and estimate the cost of the *best plan* utilizing the information given about different indexes in each part.

Q5A. Compute the estimated result size of the query, and the reduction factor of each filter.

(3 marks)

Q5B. Compute the estimated cost of the *best plan* assuming that a *clustered B+ tree* index on *quantity* is (the only index) available. Suppose there are 200 index pages.

(3 marks)

Q5C. Compute the estimated cost of the *best plan* assuming that an *unclustered Hash* index on *discount* is (the only index) available.

(2 marks)

Q5D. If you are given complete freedom to create one index to speed up this query, which index would be the best one to answer this query? Please give complete information about the index, e.g. is it clustered or unclustered, is it hash or B+ tree, which attributes will it cover.

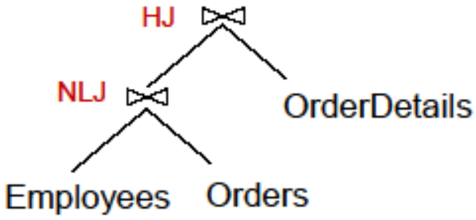
(2 marks)

Section 6 – Query Optimisation

Given the schema of Question 3 (SQL-DML), consider the relations Employees, Orders and OrderDetails. Imagine that relation Employees has 1,000 pages, relation Orders 5,000 pages, and relation OrderDetails 10,000 pages. Each page stores 100 tuples, and neither relation has any indexes built on it. Consider the following query:

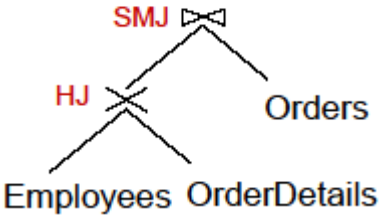
```
SELECT *
FROM Employees AS E, Orders AS O, OrderDetails AS OD
WHERE E.empid = O.empid AND O.orderid = OD.orderid;
```

Q6A. Compute the cost of the plan shown below. NLJ is a *Page-oriented* Nested Loops Join. Assume that *empid* is the candidate key of Employees, *orderid* is the candidate key of Orders, and 100 tuples of a resulting join between Employees and Orders can fit on one page.



(4 marks)

Q6B. Would the plan presented below be a valid candidate that System R would consider and compute cost for during query optimisation? Why?



(3 marks)

Q6C. Consider the query presented below. Does the following equivalence class hold? Yes/No and Why?

```
SELECT firstname, lastname
FROM Employees NATURAL JOIN Orders NATURAL JOIN OrderDetails
WHERE quantity > 5 AND freight < 100
```

$$\Pi_{\text{firstname, lastname}}(\sigma_{\text{quantity}> 5 \wedge \text{freight} <100}(\text{Employees} \bowtie \text{Orders} \bowtie \text{OrderDetails}))$$

$$\leftrightarrow \sigma_{\text{quantity}> 5 \wedge \text{freight} <100}(\Pi_{\text{firstname, lastname}}(\text{Employees} \bowtie \text{Orders} \bowtie \text{OrderDetails}))$$

(3 marks)

Section 7 – Normalisation

Q7. The table shown below is part of an office inventory database. Identify the design problems and write a revised table structure in 3rd Normal Form (3NF) that corrects those problems. For each step explicitly identify which normal form is violated and briefly explain why.

Write the normalised tables in a textual format, as in:

$\text{TableName } (\underline{\text{PrimaryKey}}, \text{ Column, } \overset{\text{FK}}{\text{ForeignKey}})$

$\text{AnotherTable } (\underline{\text{PrimaryKey}}, \text{ Column, AnotherColumn})$

Item ID is the candidate key for this table. Item ID determines Description, Quan, Cost/Unit and Dept, while Dept determines Dept Name and Dept Head.

Item ID	Description	Dept	Dept Name	Dept Head	Quan	Cost/Unit	Inventory Value
4011	5 ft desk	MK	Marketing	Jane Thompson	5	200	1000
4020	File cabinet	MK	Marketing	Jane Thompson	10	75	750
4005	Executive chair	MK	Marketing	Jane Thompson	5	100	500
4036	5 ft desk	ENG	Engineering	Ahmad Rashere	7	200	1400

(10 marks)

Section 8 – Data Warehousing

You are making a data warehouse for a real estate agency. The company wants to track information about the selling of their properties. Whenever a buyer comes into the office an agent takes that person to a number of properties and deals with the buyer. This warehouse keeps information about the agents (real estate license#, first name, last name, phone #, and branch office), buyers that come in (buyer id, first name, last name, phone #, max price), and property (property#, property address, number of rooms, pool, owner id). The information managers want to be able to find is the number of times a property is viewed, sales price and commission. Sales commission is additional compensation the agent receives for exceeding expectations. The information needs to be accessible by agent, by buyer, by property and for different time (day, week, month, quarter and year).

Q8A. Draw a *star schema* to support the design of this data warehouse, showing the attributes in each table. Use PK to denote primary key, PFK to denote primary foreign key, and FK to denote foreign key. You *do not* need to specify data types nor whether the fields are NULL/NOT NULL.

(8 marks)

Q8B. Why are star schemas preferred over relational database designs to support decision making?

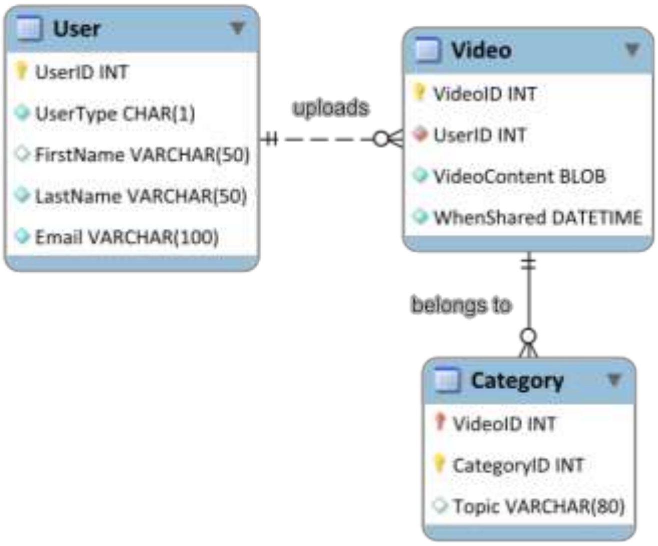
(2 marks)

Section 9 – NoSQL and Database Administration

Q9A. There are trade-offs between the principles of ACID and BASE. Discuss the trade-off between availability and consistency for relational and NoSQL databases. Illustrate your answer using either the example of Facebook or the example of Twitter.

(5 marks)

Q9B. Vine is a social media sharing service where users can host 6-second video clips within multiple categories (e.g. “Comedy”, “Science”, “Social”). Part of the database schema for the Vine service is given below.



Storage requirements for different data types	
Data type	***Storage requirement (bytes)***
INT	4
BLOB	65,535 (Max)
DATETIME	8

There are 15 different categories that users can share videos about and 1 million users to start with. A user posts 5 videos on average per month. Assume that the average storage requirement for the BLOB data type is 20,000 bytes.

Estimate the disk space requirements **only for the Video table** at go-live and after one month of operation.

(5 marks)

Section 10 – Multiple Choice Questions

There are fifteen multiple choice questions. Each question is worth 1 mark. There is only one correct answer per question. Write the question number and your answer in the script book.

Q10A. Which of the following is **NOT** part of the database development lifecycle?

A) Implementation
B) Maintenance
C) Requirement analysis
D) First-level support

Q10B. A relation which is **NOT** in the conceptual and logical models but is made available to users is a:

A) Data type
B) View
C) Revoke
D) Grant

Q10C. Which of the following is **NOT** a valid join?

A) Sort-merge
B) Hash
C) Nested loop
D) Heap

Q10D. Which of the following is **NOT** a true statement about data and information?

A) Dates and audio are two examples of data
B) Information has been processed so that it increases the audience’s knowledge
C) Information is used to create data
D) The term “data” refers to raw facts

Q10E. Which of the following is **NOT** true about normal forms?

A) A table in 1NF must have no multivalued attributes
B) A table in 2NF must have no multivalued attributes
C) A table in 2NF must have no transitive functional dependencies
D) A table in 3NF must have no transitive functional dependencies

Q10F. Which of the following is **NOT** one of the basic operations of Relational Algebra?

A) Set-difference
B) Cross-product
C) Equality
D) Union

Q10G. The structure of a Clustered B+ tree Index does **NOT** contain:

A) Data pages
B) Leaf nodes
C) Internal nodes
D) Root node

Q10H. Which of the following is **NOT** a true statement about projection?

A) Projection can be accomplished by hashing
B) The cost of projection by external merge sort depends on the reduction factor
C) Projection algorithms are designed to remove duplicates from the result set
D) Projection by external merge sort uses the same sorting approach as sort-merge join

Q10I. Which of the following is **NOT** a commonly-used level of lock granularity?

A) Field-level locking
B) Row-level locking
C) Table-level locking
D) Database-level locking

Q10J. Which of the following is **NOT** a valid type of logical database backup?

A) Incremental backup
B) Online backup
C) Onsite backup
D) Offline backup

Q10K. Which of the following is **NOT** a normalization concept?

A) Non-key dependency
B) Partial functional dependency
C) Candidate key
D) Determinants

Q10L. Which of the following is **NOT** a file organisation type?

A) Hash file organisation
B) Index file organisation
C) Heap file organisation
D) Sorted file organisation

Q10M. Which of the following is **NOT** an SQL command category?

A) DDL
B) DML
C) DCL
D) DSL

Q10N. Which of the following is **NOT** a part of query optimisation?

A) Plan enumeration
B) Costing
C) Building a new index
D) Result size estimation

Q10O. Which of the following is **NOT** true about choosing data types?

A) May help improve query optimisation
B) Help DBMS store and use information efficiently
C) Help minimise storage space
D) May help improve data integrity

END OF THE EXAM