Victoria University of Wellington

DEGREE EXAMINATIONS — 1995

(MID-YEAR)

COMP 302

Database Systems

Time Allowed: 3 hours.

Instructions: Answer all questions.
Make sure your answers are clear and to the point.
No reference material is allowed.
There are 180 possible marks on the exam.

1. Basic Concepts

(a) A file is a sequence of records. Give one reason why the allocation of records to blocks is an important issue in database system performance.
If different records in a file have different sizes, the file is said to be made up of variable length records. List two reasons for having variable-length records.
[3 marks]

(b) Consider the schema of an object-oriented database and the record structures of the **department** and **student** records where the **department** records are physically grouped together and the **student** records are physically grouped together.
The most common query is to retrieve a **department** record and all records for **student**s majoring in that department.
Describe a file organization that makes the processing of this query more efficient.
Would this organization require any changes to the schema? If so, what? If not, why not?
Would this require any changes to the structure of the record? If so, what? If not, why not?
[6 marks]

(c) Clustering indices may allow faster access to data than a nonclustering index. Why? When is it necessary to create a nonclustering index despite the advantages of a clustering index?
[3 marks]

(d) A file has r = 20,000 **student** records of fixed length. Each record has the following fields: **name** (30 bytes), **ssn** (9 bytes), **address** (40 bytes), **phone** (9 bytes), **birthdate** (8 bytes), **sex** (1 byte), **majordeptcode** (4 bytes), **minordeptcode** (4 bytes), **classcode** (4 bytes, integer) and **degreeprogram** (3 bytes). An additional byte is used as a deletion marker. The file is stored on a disk with the following characteristics: block size B = 512 bytes, interblock gap size G = 128 bytes, block pointer P = 6 bytes, record pointer = 7 bytes, number of blocks per track = 20, number of tracks per surface = 400. A disk pack consists of 15 double-sided disks. (In the following questions, an expression using the data above that represents the answer will suffice.)

i. Calculate the record size R in bytes.

[1 mark]

ii. Calculate the blocking factor **bfr** and the number of file blocks **b**, assuming an unspanned organization.

[2 marks]

iii. Calculate the average time it takes to find a record by doing a linear search of the file if the records are not stored contiguously.

[2 marks]

iv. Assume that the file is ordered by **ssn**; calculate the time it takes to search for a record given its ssn value, by doing a binary search.

[2 marks]

v. Suppose again that the file is ordered by the key field **ssn** and we want to construct a *primary index* on **ssn**. Calculate

A. the index blocking factor bfr_i (which is also the index fan-out);

[1 mark]

B. the number of first-level index entries and the number of first-level index blocks;

[1 mark]

C. the number of levels needed if we make it into a multi-level index;

[1 mark]

D. the total number of blocks required by the multi-level index; and

[3 marks]

E. the number of block accesses needed to search for and retrieve a record from the file - given its **ssn** value - using the primary index.

[1 mark]

vi. Suppose that the file is not ordered by the key field **ssn** and we want to construct a B+ tree access structure on **ssn**. Calculate

A. the orders p and p_{leaf} of the B+ tree;

[2 marks]

B. the number of leaf-level blocks needed if the blocks are approximately 69% full (rounded up for convenience);

[2 marks]

C. the number of levels needed if internal nodes are also 69% full (rounded up for convenience);

[4 marks]

D. the total number of blocks required by the B+ tree; and

[1 mark]

E. the number of block accesses needed to search for and retrieve a record from the file - given its **ssn** value - using the B+ tree.

[1 mark]

2. Relational Model

(a) Give one reason why the relational model is popular. List two reasons why more advanced models, like the object-oriented data model, are being investigated. Can a query language like SQL be extended to include the concepts introduced by the more advanced models? If so, explain how two of the concepts may be included in SQL? If not, why not?

[6 marks]

(b) Consider the following relational database:

lives(person_name, street, city)
works(person_name, company_name, salary)
located_in(company_name, city)
manages(person_name, manager_name)

Give an expression in:

- relational algebra
- tuple relational calculus
- domain relational calculus

for each of the queries below:

i. Find the name and city of all the employees who work for BNZ.

[6 marks]

ii. Find all the cities that exist in both the **lives** relation and the **located_in** relation.

[6 marks]

iii. Find the names of all employees who do not work for BNZ (without using $\neq$ in the condition).

[6 marks]

(c) Construct an E-R diagram for a car insurance company with a set of customers, each of whom owns a number of cars. The kind of information we would be interested in is the name, social security number and adress of each customer, the year, model, registration number and a record of the accidents associated with each car, and the driver, date, and dollar amount of the damage of each accident.

[6 marks]

(d) Construct appropriate tables (or relations) for the E-R diagram in part (c). Indicate the keys and state any assumptions you make.

[6 marks]

3. Database Design

(a) Attributes can be grouped to form a relational schema by using common sense or by mapping a schema specified using the ER model into a relational schema. What concept can be used to measure the quality of the reulting schema?

[1 mark]

(b) What does a functional dependency specify?

[2 marks]

Give an example of a functional dependency. Explain this example.

[3 marks]

(c) Give expressions to represent the following functional dependencies:

i. A one-to-one relationship set exists between entity sets *student* and *advisor*.

[2 marks]

ii. A many-to-one relationship set exists between entity sets *student* and *advisor*.

[2 marks]

(d) Give one reason why in designing a relational database we might choose a non-BCNF design?

[1 mark]

(e) Consider the universal relation $R = \{A, B, C, D, E, F, G, H, I\}$ and the set of functional dependencies $F = \{\{A, B\} \to C, A \to \{D, E\}, B \to F, F \to \{G, H\}, D \to \{I, J\}\}$.

i. What is the key for R?

[3 marks]

ii. Decompose R into 2NF.

[5 marks]

iii. Decompose R into 3NF.

[5 marks]

iv. Consider the following decompositions for the relational scheme R. Determine whether each decomposition has

A. the dependency preservation property, and

B. the lossless join property

with respect to F. Also determine which normal form each relation in the decomposition is in.

A. $D_1 = \{R_1, R_2, R_3, R_4, R_5\}$; $R_1 = \{A, B, C\}$, $R_2 = \{A, D, E\}$, $R_3 = \{B, F\}$, $R_4 = \{F, G, H\}$, $R_5 = \{D, I, J\}$.

[6 marks]

B. $D_3 = \{R_1, R_2, R_3, R_4, R_5\}$; $R_1 = \{A, B, C, D\}$, $R_2 = \{D, E\}$, $R_3 = \{B, F\}$, $R_4 = \{F, G, H\}$, $R_5 = \{D, I, J\}$.

[6 marks]

4. Query Processing and Optimization

(a) What are the two main approaches for query optimization? Give a general description of each approach.

[4 marks]

Now consider the object-oriented data model. Could the same approaches be used in this setting? Explain your answer.

[6 marks]

(b) Consider the following SQL query.

select emp.fname, emp.lname, emp.address
from employee emp, department dept
where dept.dname = 'Research' and dept.dnumber = emp.dno

i. Draw two query graphs that can represent this query.

[4 marks]

ii. Under what circumstances would you use each of your graphs?

[4 marks]

(c) Explain the heuristics for algebraic optimization. Illustrate the heuristics with an example.

[10 marks]

(d) Consider a **project** operation $\pi_{\langle attribute\ list \rangle}(R)$. If $\langle attribute\ list \rangle$ does not include a key of relation R, then duplicate tuples must be eliminated. This can be done by sorting the result of the operation and then eliminating duplicate tuples, which appear consecutively after sorting. Develop a cost function for this **project** operation. (Remember to include the cost of reading and writing records to and from disk.)

[8 marks]

5. Transactions, Concurrency and Recovery

(a) What does a transaction represent? Describe the four properties that are considered desirable in transactions.

[6 marks]

(b) Explain the wait-die and wound-wait protocols, deadlock and livelock. Using an example, show how the wait-die and wound-wait protocols avoid deadlock and livelock.

[10 marks]

(c) Describe cascading roll-back. Give an example illustrating cascading rollback. Why are most recovery mechanisms designed so that cascading rollback is never required?

[5 marks]

(d) A single transaction may require access to multiple databases on different machines. List five issues that arise when a transaction runs on multiple databases that don't arise when a transaction runs on a single database.

[5 marks]

(e) Describe a recovery mechanism for a system where a transaction may access multiple databases.

[10 marks]

****