University of Glasgow

Wednesday 11 May 2022
14:00-15:30 BST
Duration: 1 hour 30 minutes
Additional time: 30 minutes
Timed exam – fixed start time

DEGREES of MSci, MEng, BEng, BSc, MA and MA (Social Sciences)

Database Systems H
COMPSCI4013

Answer All Questions

This examination paper is an open book, online assessment and is worth a total of 60 marks.

Part A: Relational Modelling and SQL [20 Marks]

Question 1.

Consider the following relational schema:

MANUFACTURER(ID, Name)

VEHICLE(NumberPlate, Price, Year, MID)

where, a manufacturer has a unique identifier (ID) and a name, while, a vehicle has a unique number plate, a price (£), a year of production and is associated with a manufacturer. The MID attribute in the Vehicle relation is a foreign key referencing to the ID primary key attribute in the Manufacturer relation. A manufacturer can produce more than one car. The primary keys are underlined.

(a) For each manufacturer, show the price of its most expensive car(s) *using* the GROUP BY clause in your SQL query. It is possible that more than one car has the same price.

[5 Marks]

(b) For each manufacturer, show the price of its most expensive car(s) *without using* the GROUP BY clause in your SQL query. It is possible that more than one car has the same price.

[10 Marks]

(c) From each manufacturer, which has produced more than 1000 cars, how many of these cars have a price greater than £100,000.

[5 Marks]

Part B: File Organization & Indexing [20 Marks]

Question 2. Consider the relation EMPLOYEE(SSN, Surname, Salary, Age, DNO), where SSN is the social security number and DNO is the department number. Consider the following context:

- The number of the distinct values of DNO is n = 100. The DNO values are the integers {1, 2, …, 100}. The DNO values are uniformly distributed across the employee tuples.
- The relation EMPLOYEE has r = 10,000 tuples, the size of each record is R = 250 bytes (DNO = 50 bytes, Salary = 100 bytes, Age = 50 bytes, SSN = 50 bytes), the block size is B = 512 bytes, and any pointer has size V = 10 bytes.
- The file of the relation EMPLOYEE is sorted by the non-key, ordering attribute DNO.
- The SQL1 query below fetches those employees working in the departments 10, 11, …, 29.

```
SQL1: SELECT * FROM EMPLOYEE WHERE DNO >= 10 AND DNO <= 29
```

(a). Calculate the number of the block accesses for SQL1 using a linear search with *exiting feature* over the ordering, uniformly distributed, non-key DNO attribute.

[5 Marks]

(b). Create a *multi-level* Clustering Index over the DNO attribute and calculate the number of the block accesses for SQL1.

[10 Marks]

(c). Consider the following SQL2 query for a specific DNO value ***x***:

```
SQL2: SELECT * FROM EMPLOYEE WHERE DNO = x
```

Find the maximum DNO value ***x*** such that searching using the linear search with exiting feature requires a smaller number of block accesses than using the Clustering index in Q2(b). Explain your answer.

[5 Marks]

Part C: Query Processing and Optimization [20 Marks]

Question 3. Consider the relations EMPLOYEE(SSN, SURNAME, AGE) and DEPENDENT(ESSN, NAME) with r_E = 1,000 tuples in Employee and r_D = 500 tuples in Dependent. Each attribute in the EMPLOYEE has size 50 bytes. Each attribute in the DEPENDENT has size 50 bytes with NDV(ESSN) = 50 distinct social security numbers (NDV stands for Number of Distinct Values) in DEPENDENT, and NDV(AGE) = 100 distinct age values in EMPLOYEE. The ESSN values are uniformly distributed across the DEPENDENT tuples. The AGE values are uniformly distributed across the EMPLOYEE tuples and range from 25 (inclusive) to 65 (inclusive). The size of the block is B = 256 bytes and any pointer has size V = 50 bytes. Consider the following query:

```
SELECT *
FROM   EMPLOYEE AS E, DEPENDENT AS D
WHERE  E.SSN = D.ESSN AND E.AGE >= 45
```

and consider the following available access paths:

- Clustering Index over the AGE in the relation EMPLOYEE.
- Clustering Index over the ESSN in the relation DEPENDENT.

(a). Calculate the number of entries and number of blocks of the two Clustering Indexes.

[5 Marks]

(b). Provide **two** *processing methods* using the Clustering Indexes from Question 3(a) and select the *best* one in terms of block accesses.

[15 Marks]

The allocated memory of the Database system is 5,000 blocks.

Note 1: For convenience in calculations, $\log_2(50) = 5.64$ and $\log_2(25) = 4.64$

Note 2: The indexes are **not** multi-level Clustering indexes.