THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

This exam paper must not be removed from the venue

Venue ____________________

Seat Number ________

Student Number |__|__|__|__|__|__|__|__|

Family Name _____________________

First Name _____________________

School of Information Technology and Electrical Engineering

EXAMINATION

Semester Two Final Examinations, 2020

INFS3200/7907 Advanced Database Systems

This paper is for St Lucia Campus students.

Examination Duration: 120 minutes

Reading Time: 10 minutes

Exam Conditions:

This is a Central Examination

This is a Closed Book Examination - no materials permitted

During reading time - write only on the rough paper provided

This examination paper will be released to the Library

Materials Permitted In The Exam Venue:

(No electronic aids are permitted e.g. laptops, phones)

Calculators - Casio FX82 series or UQ approved (labelled)

Materials To Be Supplied To Students:

None

Instructions To Students:

Additional exam materials (eg. answer booklets, rough paper) will be provided upon request.

Please answer all questions on the examination paper.

Total is 60 marks.

For Examiner Use Only

Question	Mark

Total ________

Question 1 [10 marks] In a distributed database system, relations are usually fragmented and distributed to multiple sites to enable parallel execution of queries.

(a) [3 marks] When performing fragmentation on a relation, what properties should the fragmentation meet to ensure the correctness? List their names and meanings.

(b) [3 marks] Given a relation R(A, B) and two simple predicates (P1: R.A $\geqslant$ 75, P2: R.A < 50), generate a set of minterm predicates which satisfy the aforementioned properties.

(c) [4 marks] Consider two relations R(A, B) and S(X, A, C), where S.A is the foreign key point to R.A. Assume that the relation R(A, B) is located on site 1 and the relation S(X, A, C) is located on site 2. Consider a join query R $\bowtie_A$ S issued at site 1. Please give a step-by-step query execution plan using semijoin operations to process this query (using either plain English or SQL queries).

Question 2 [6 marks] Data replication is very important in distributed database design.

(a) [2 marks] List four types of replication strategies and point out whether they are synchronous replication or asynchronous replication.

(b) [2 marks] Describe what needs to be maintained/satisfied for a voting-based approach to guarantee the data consistency among replications.

(c) [2 marks] If a database is read-intensive with rare updates, should we use a bigger number of write copies in the voting-based approach? Why or why not?

Question 3 [4 marks] Answer the following questions:

(a) [2 marks] A data warehouse is usually represented by either a star schema or a snowflake schema. What are the advantages and disadvantages of a star schema compared with a snowflake schema?

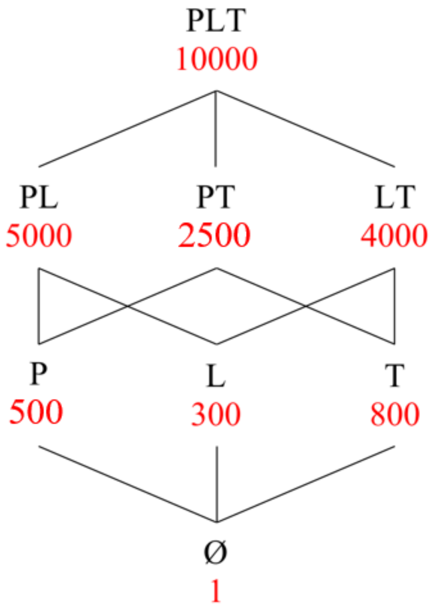
(b) [2 marks] What is a weak identifier? How can we deal with weak identifiers when converting a star schema to a snowflake schema? Explain both questions by either definitions or examples.

Question 4 [12 marks] Materialized cuboids are pre-computed and stored on the disk. A data warehouse can often make use of materialized cuboids.

(a) [4 marks] In general, for a 4-dimensional data cube with no hierarchical structure, how many materialized cuboids can be pre-computed? What if two of the dimensions contains two levels (e.g. *product* < *category* for *product* dimension)?

(b) [2 marks] Given the base cuboid on {*product*, *location*, *time*}, how can we obtain the total sale of the "Dell Laptop" product at each location using OLAP operations? (**Hint**: you only need to explain which operations are performed on which dimensions in either SQL or plain English)

(c) [6 marks] Given the above cube with three dimensions (P, L, T) and one measure *sales*. Below is a lattice of all possible cuboids created on the data warehouse. Each of the numbers shows the cost of using the corresponding cuboid when it is materialized to answer a group-by query.



Assume that the frequency distribution of all group-by queries is as follows:

{PTL (0.05), PL (0.1), PT (0.05), LT (0.1), P (0.1), L (0.3), T (0.15), Ø (0.15)}

What are the first three cuboids that should be materialized in order to minimize the total query cost and why? Please provide the cost calculation of every cuboid to justify your answer.

Question 5 [9 marks] Data integration is an important pre-processing step in data warehousing and data mining.

(a) [4 marks] List at least four challenges we need to address in data integration. Give one example for each challenge.

(b) [1 mark] Consider the following two University data models. University A stores staff records in one table:

Staff(Emp#, Fname, Lname, Bdate, Dept#)

University B stores staff records in two tables for departments 01 and 02 separately:

Dept_01(Eid, Fname, Sname, Position, Phone#)
Dept_02(Eid, Fname, Sname, Position, Phone#)

It is known that Lname matches Sname, Fname matches Fname, and Emp# matches Eid. Define a global schema we can construct from these data models.

(c) [4 marks] Please write an SQL query to generate the global schema. (**Hint**: using views)

Question 6 [8 marks] Answer the following questions:

(a) [2 marks] The similarity between two strings can be measured by Edit distance, for two strings with m and n characters respectively, what is the maximum possible edit distance?

(b) [1 mark] In Edit distance, what operation is considered as an 'edit'/'transformation'?

(c) [5 marks] The string similarity can also be measured by the Jaccard coefficient based on q-grams. Compute the Jaccard coefficient between two strings "world peace" and "peace world" when q=4. Without calculating their Edit distance, which measurement do you think better reflects their actual similarity? Why?

Question 7 [6 marks] Data privacy is a very important issue when publishing data. Various techniques are introduced to avoid the leak of personal information when data are published.

(a) [2 marks] K-anonymity is a common and simple solution to privacy-preserving data publishing. What is the meaning of k-anonymity?

(b) [2 marks] What possible problems may happen in k-anonymity? Describe it by either explanations or examples, and propose a solution if possible.

(c) [2 marks] What is the main feature of differential privacy? Specify at least one key difference between differential privacy and k-anonymity in terms of the design rationale.

Question 8 [5 marks] A relational database system is not suitable to manage spatial data nor to process spatial queries. For a set of polygons where each polygon is defined using a sequence of points, and a point is represented using its two coordinates X and Y.

Using a relational database system, one can design a table Polygon (polygonID, pointIndex, X, Y) to each record, for example, (10, 3, 12.5, 24.6) in the table means the 3rd point of polygon 10 is point (12.5, 24.6).

(a) [2 marks] For a given polygon ID (say, 12345), how can we retrieve this polygon from the table? The points in the retrieved polygon should be ordered based on their pointIndex numbers. Please use SQL to write such a query.

(b) [2 marks] With your design above, how do you find all polygons that contain a query point q? Assume that we have a function contains (polygon p, point q) that returns true if point q is inside polygon p.

(c) [1 mark] How an R-tree index can make the execution of the query above efficiently?

END OF EXAMINATION