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, 2019

INFS3200/INFS7907 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 50 marks.

For Examiner Use Only

Question	Mark

Total ________

Question 1. Distributed Databases (8 marks)

A Semijoin is a special type of join operation that can be used in distributed database design and distributed query processing.

(a) [4 marks] Consider two relations R(A, B) and S(X, A, C), where S.A is the foreign key. Assume that R is horizontally fragmented based on its attribute A into R1 and R2. Please use the semijoin operation to define the derived horizontal fragmentation of S based on the fragmentation of R, and explain how your S fragmentation meets the reconstruction property.

(b) [4 marks] Assume that the relation R(A, B) is located on site 1 and that the relation S(X, A, C) is located on site 2. Consider a join query R $\bowtie_A$ S at site 1. Please give a step-by-step query execution plan using semijoin operations to process this query.

Question 2. Data warehouses (11 marks)

Consider a sales fact table with three dimensions (time, location, product).

(a) (3 marks) Explain what a data cube is in data warehousing systems.

(b) (3 marks) Explain what a dicing operation is.

(c) (2 marks) It is not common for data warehousing systems to support update operations. Describe a reason why supporting updates in data warehouses is not a good idea. Briefly justify your answer.

(d) (3 marks) A data warehouse can often make use of materialized views (e.g., using materialized data cubes). Discuss advantages and disadvantages of building materialized views in data warehouses.

Question 3. Data Warehouses (6 marks)

Materialized cuboids are pre-computed and stored on disk. A data warehouse can often make use of materialized cuboids.

a) (2 marks) Suppose that the cuboid on *{student, semester}* is materialized. Among the following group-by queries, which queries can benefit from this materialized cuboid?

{student, course, semester}

{student, course}

{student, semester}

{course, semester}

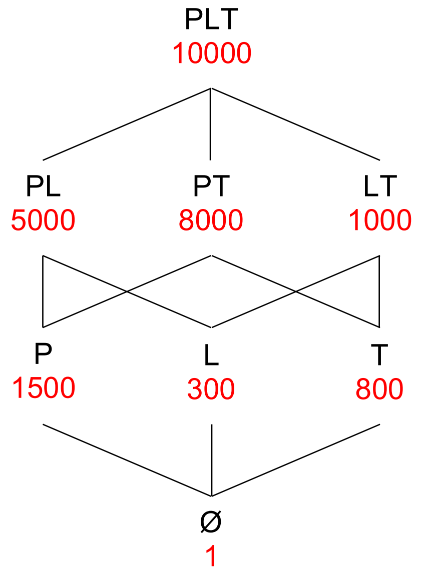
{student}

{course}

{semester}

Ø

b) (4 marks) Suppose that a data warehouse consists of the following three dimensions: *product* (P), *location* (L), and *time* (T), and one measure *sales*. Below is a lattice of all possible cuboids created on the data warehouse dimensions. Each of the numbers shows the cost of using the corresponding cuboid when it is materialized to answer a group-by query. Assume that all the queries are issued with the same frequency, and we have already materialized two cuboids: *{PLT}* and *{PL}*. Which cuboid should be materialized next using the greedy algorithm and why?



Question 4. Data Integration (9 marks)

(a) (3 marks) For two strings with *m* and *n* characters respectively, which is the maximum possible edit distance?

(b) (3 marks) What is the edit distance between "maple" and "apple"? Please show the matrix of your calculation.

(c) (3 marks) String similarity can also be measured using Jaccard coefficient based on q-grams. It is a more suitable string similarity measure than the edit distance for two strings that have words in different orders, such as "CEO of Apple" versus "Apple CEO". Why?

Question 5. Modern Platforms (6 marks)

(a) [3 marks] Explain the main limitation of the Google File System design.

(b) [3 marks] Explain the main efficiency bottleneck of Map/Reduce.

Question 6. Privacy (10 marks)

K-anonymity and differential privacy are two common solutions to privacy-preserving data publishing. For each of these two solutions, please explain (1) what they mean, and (2) what changes they need to make to the data before publishing.

(a) [5 marks] K-anonymity.
(b) [5 marks] Differential privacy.

END OF EXAMINATION