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 One Final Examinations, 2019

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.

50 marks in total.

For Examiner Use Only

Question	Mark

Total ________

Question 1 [6 marks]. Data replication is very important in distributed database design.

(a) [2 marks] What are the benefits of having data replications, and at what costs?

(b) [2 marks] Describe how a voting-based approach works to maintain data consistency among data replications.

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

Question 2 [5 marks]. Consider a simplified database defined by the following schemas:

STUDENT(SNO, SNAME, PROGRAM)

COURSE(CNO, CNAME, CTITLE)

ENROLLMENT(SNO, CNO, RESULT)

(a) [1 mark] Given the following SQL query, transform it into a query execution tree.

SELECT SNO, SNAME, CNAME, RESULT

FROM STUDENT S, COURSE C, ENROLLMENT E

WHERE C.CNO = E.CNO AND S.SNO = E.SNO

(b) [2 marks] Assume that relation COURSE is horizontally fragmented as follows:

$\text{COURSE}_1 = \sigma_{\text{CNO} \leqslant \text{C100}}(\text{COURSE})$

$\text{COURSE}_2 = \sigma_{\text{CNO} > \text{C100}}(\text{COURSE})$

and that relation ENROLLMENT is horizontally fragmented as follows:

$\text{ENROLLMENT}_1 = \sigma_{\text{CNO} \leqslant \text{C100}}(\text{ENROLLMENT})$

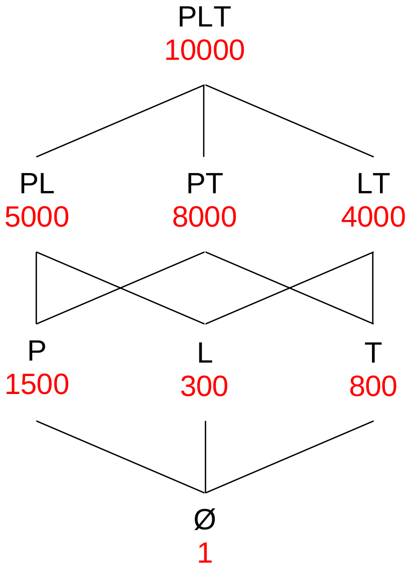
$\text{ENROLLMENT}_2 = \sigma_{\text{C100} < \text{CNO} \leqslant \text{C200}}(\text{ENROLLMENT})$

$\text{ENROLLMENT}_3 = \sigma_{\text{CNO} > \text{C200}}(\text{ENROLLMENT})$

Show an equivalent query execution tree of the above query after replacing relations with fragments.

(c) [2 marks] Show an optimised query execution tree of the above query after applying reduction rules.

Question 3 [8 marks]. Suppose that a data warehouse for *Company* 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. Each of the numbers shows the cost of using the corresponding cuboid when it is materialized to answer a group-by query.



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

{PTL (0.05), PL (0.25), PT (0.15), LT (0.1), P (0.2), L (0.1), T (0.1), Ø (0.05)}

What are the first two cuboids that should be materialized in order to minimize total query cost, and why?

Question 4 [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, and give one example for each challenge.

(b) [1 mark] Consider the following two University data models:

University A stores student records in one table:

Student(S#, Fname, Lname, Bdate, Program#)

University B stores student records in two tables for programs 01 and 02 separately:

Prog_01(Sid, Fname, Sname, Credit, email)

Prog_02(Sid, Fname, Sname, Credit, email)

It is known that Lname matches Sname, and S# matches Sid. Define the global schema we can construct from these data models.

(c) [4 marks] Write a SQL query to generate the global schema. (Hint: using views)

Question 5 [15 marks]. Data quality issues need to be addressed before the data can be released for use by other data analysis applications.

(a) [4 marks] Data quality can be measured from various dimensions. Please list at least four data quality dimensions and give one example of data quality problem for each of these dimensions.

(b) [1 mark] Record linkage is an important task in data quality management. Explain the meaning of record linkage.

(c) [4 marks] Edit distance is a common string similarity measure used in record linkage. Edit distance between two strings is the minimum number of operations (i.e., insert, delete, or replace one character) to transform one string to the other. Compute edit distance between two strings "Serious" and "Ceriers" using the dynamic programming algorithm. Show the calculation step by step in a matrix. What is the edit distance between these two strings?

(d) [2 marks] Jaccard coefficient is another string similarity measure that can be used for record linkage. Assume that we need to use either edit distance or Jaccard coefficient to perform record linkage for a dataset of people's names. Which similarity measure do you suggest to use in the following cases respectively, and why?

- Names are written as either *{first name, last name}* or *{last name, first name}.*
- All the names are written as *{first name, last name}*, but they contain some minor typos.

(e) [4 marks] Efficiency of record linkage should also be considered in practice. Various techniques have been proposed to reduce the number of record comparisons, such as Blocking, Sorted Neighbourhood Approach, Clustering and Canopies, etc. Please explain one of these techniques.

Question 6 [7 marks]. Data privacy is a very important issue when publishing data. K-anonymity is a common and simple solution to privacy-preserving data publishing.

(a) [1 mark] What is K-anonymity?

(b) [3 marks] Describe the general approach of K-anonymity.

(c) [1 mark] K-anonymity is still vulnerable in some situations. Explain possible problems of K-anonymity.

(d) [2 marks] L-diversity is a method to reduce the vulnerability of K-anonymity. Describe the general approach of L-diversity, especially its difference with K-anonymity.

END OF EXAMINATION