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, 2017

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:

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 [5 marks] Consider relation R(A, B, C, D), where A is the primary key attribute. R is vertically fragmented into R1(A, B) and R2(A, C, D) and allocated at site N1 and N2 respectively.

(a) [3 marks] How to insert a tuple (a, b, c, d) into R? This insert operation must meet the atomicity property.

(b) [2 marks] Construct a simple example to illustrate the problems if attribute A does not appear in both fragments, for example, R is fragmented into R1(A, B) and R2(C, D).

Question 2 [10 marks] A webpage can be uniquely identified by its URL. There can be hyperlinks among Webpages. If there is a hyperlink from webpage X to webpage Y, we say X points to Y. Many web search ranking algorithms need to find, for each webpage X, all webpages which X points to (called out-going webpages of X), and all webpages which point to X (called in-coming webpages of X).

(a) [8 marks] One team needs to use MapReduce to compute the number of in-coming webpages for all webpages that are already stored in GFS. Please provide properly commented pseudo code for map() and reduce() functions to perform this task. You can use any programming language or just plain English.

(b) [1 mark] There can be many nodes to execute the map() and reduce() functions. What does MapReduce system do between the map phase and reduce phase?

(c) [1 mark] What happens if a node executing map() or reduce() function fails during execution?

Question 3 [8 marks] Suppose that a data warehouse for *University* consists of the following three dimensions: *student*, *course*, and *semester*, and one measure *grade*. At the lowest conceptual level (e.g., for a given student, course, and semester combination), the *grade* measure stores the actual course grade of the student. At higher conceptual levels (e.g., for a given student and course combination), *grade* stores the average grade for the given combination.

(a) [4 marks] The data warehouse can be modelled by either a *star schema* or a *snowflake schema*. Briefly describe the similarities and the differences between the two models, and then analyse their advantages compared to one another.

(b) [2 marks] Given the base cuboid on *{student, course, semester},* how can we obtain the average grade of "Advanced Database Systems" course for each student using OLAP operations? (*Hint: you only need to explain which operations are performed on which dimensions in either SQL or plain English*)

(c) [2 marks] *Bitmap indexing* is a useful technique in data warehousing. Taking this cube as an example, briefly discuss the advantages and problems of using a bitmap index structure.

Question 4 [10 marks] *Materialized cuboids* are pre-computed and stored on the disk. A data warehouse can often make use of materialized cuboids.

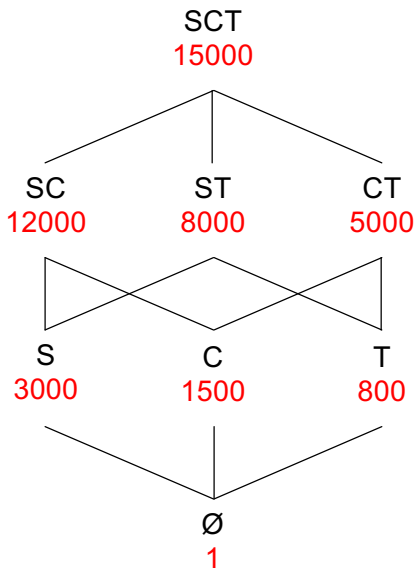
(a) [2 marks] Discuss the advantages and disadvantages of building materialized cuboids in a data warehouse.

(b) [2 marks] Consider the data warehouse defined in Question 3. If each dimension has four levels (e.g., "student < major < school < university" for *student* dimension), how many cuboids will this cube contain?

(c) [2 marks] Suppose that the cuboid on *{student, course}* 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

(d) [4 marks] Below is a lattice of all possible cuboids created on the data warehouse of Question 3. We ignore dimension hierarchies. S, C, and T represent *student*, *course*, and *semester*, respectively. Each of the numbers is the cost of using the corresponding cuboid when it is materialized to answer a group-by query. Suppose that all the queries are issued with the same frequency. Which two cuboids will be materialized first using the *greedy algorithm* and why?



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, and 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 the 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] Entity resolution plays an important role in data integration. *Edit distance* and *Jaccard coefficient* are two common string similarity measures used in entity resolution.

(a) [2 marks] Edit distance between two strings is the *minimum* number of operations (i.e., insert, delete, or replace one character) to transform one string to another. Given two strings with *m* and *n* characters respectively, what are the minimum and maximum possible edit distances?

(b) [2 marks] For a string of *n* characters, how many *q*-grams does it contain? (Assume no # symbols will be added at the beginning or the end of the string when generating *q*-grams).

(c) [2 marks] Suppose that we need to use edit distance or Jaccard coefficient to perform entity resolution 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.

(d) [2 marks] Please give an example that using string similarity alone cannot solve the problem of entity resolution.

Question 7 [10 marks] Data privacy is a very important issue when publishing data. 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