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, 2018

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 60 marks.

For Examiner Use Only

Question	Mark

Total _____

Question 1. Distributed Databases (3 marks)

Consider relation S(A, B, C, D), where A is the primary key attribute. S is vertically fragmented into S1(A, B) and S2(A, C, D) and allocated at site N1 and N2 respectively.

How to insert a tuple (a, b, c, d) into S? This insert operation must meet the atomicity property.

Question 2. Distributed Databases (4 marks)

When a relational table is vertically partitioned into two tables, the primary key needs to be duplicated in both partitions. Please construct a simple example to illustrate the problem(s) that can be caused if the primary key is not duplicated.

Question 3. Data warehouses (13 marks)

(1) (3 marks) Explain what a data cube is in data warehousing systems.

(2) (3 marks) Explain what a slicing operation is.

(3) (3 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.

(4) (4 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.

(Additional writing space for Question 3.)

Question 4. Data Warehouses (10 marks)

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

(1) (4 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}

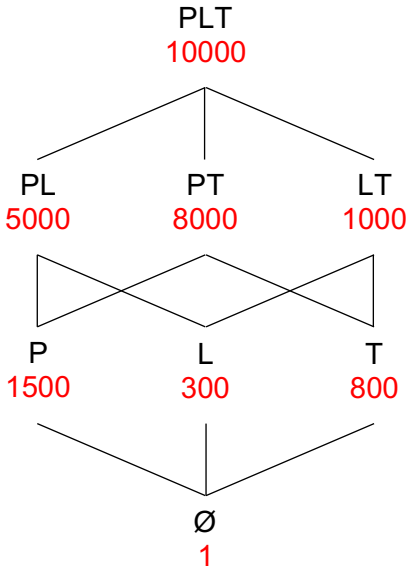
{student}

{course}

{semester}

Ø

(2) (6 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. Assume that all the queries are issued with the same frequency, and we have already materialized two cuboids {*PLT*} and {*LT*}. Which cuboid will be materialized next using the greedy algorithm and why?



(Additional writing space for Question 4.)

Question 5. Data Integration (10 marks)

(1) (2 marks) For two strings with *m* and *n* characters respectively, what is the minimum possible edit distance?

(2) (3 marks) What is the edit distance between “iphone” and “ifone”? Please show the matrix of your calculation.

(3) (5 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 “President of the USA” versus “USA President”. Explain why it is more suitable and compute the similarity between the two strings for q=3.

(Additional writing space for Question 5.)

Question 6. Knowledge Graphs (10 marks)

(1) (5 marks) Draw an RDF Schema or OWL diagram to represent movie information. Important classes will include: Actor (each having a name and a date of birth), Stars (i.e., famous Actors), Producer, Movie, Award, Genre (e.g., Documentary, Fiction) and Sub-Genre (e.g., Science-Fiction, Western, etc. for a Fiction). Please include as many subclasses and constraints (e.g., domain or range) as possible.

(2) (3 marks) Draw a small RDF instance of your schema, including one movie and a couple of actors.

(3) (2 marks) For the following SPARQL query and RDF triples, what is the query result?

Query

```
PREFIX gr: <http://purl.org/goodrelations/v1#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?x
WHERE { ?y gr:UnitPriceSpecification ?price .
        FILTER (?price < 50000) .
        ?y gr:hasBrand ?z .
        ?z rdfs:Label ?x
}
```

Data

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix gr: <http://purl.org/goodrelations/v1#> .
@prefix car: <http://example.org/car/> .
@prefix brand: <http://example.org/brand/> .
car:c1 gr:hasBrand brand:b1.
car:c1 gr:ProductOrServiceModel “Mustang GT” .
car:c1 gr:UnitPriceSpecification 52000 .
car:c2 gr:hasBrand brand:b2 .
car:c2 gr:ProductOrServiceModel “Triton GLX” .
car:c2 gr:UnitPriceSpecification 42000 .
brand:b1 rdf:Type gr:Brand .
brand:b1 rdfs:Label “Ford” .
brand:b2 rdf:Type gr:Brand .
brand:b2 rdfs:Label “Mitsubishi” .
```

(Additional writing space for Question 6.)

Question 7. Privacy (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