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, 2018

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 [11 marks] Data fragmentation and data replication are two important steps of distributed database design.

(a) [4 marks] Suppose that we have the following two relations *MachineType* and *Skills*, where the attribute *Machine* in relation *Skills* is a foreign key referring to the attribute *Machine* in relation *MachineType*. Given the following two predicates, what are the primary horizontal fragments for relation *MachineType* and the derived horizontal fragments for relation *Skills*? (Hint: the answer should contain four tables in total)

- Type = “CPU”
- Type = “Tablet PC”

Skills

Person	**Machine**
Lisa	AMD A4-3300
John	AMD A4-3300
Brown	AMD A4-3300
John	Samsung700T1A-A01
Young	Samsung700T1A-A01
Kate	Toshiba AT100
Brown	HL-3070CW
John	HL-3070CW

MachineType

Type	**Machine**
CPU	AMD A4-3300
CPU	Intel Core i7
Tablet PC	Samsung700T1A-A01
Tablet PC	Samsung700T1A-H01
Tablet PC	Toshiba AT100
Printer	HL-3070CW

(b) [2 marks] The correctness of a fragmentation is usually measured by three criteria, namely completeness, disjointness, and reconstructability. Is the above fragmentation correct, why or why not?

(c) [3 marks] What is data replication? What are the benefits of having data replications, at what costs?

(d) [2 marks] Voting-based approach is often used in a distributed system to maintain data consistency among data replications. Please explain how such a technique works to manage read and write operations on a data object that has *N* copies.

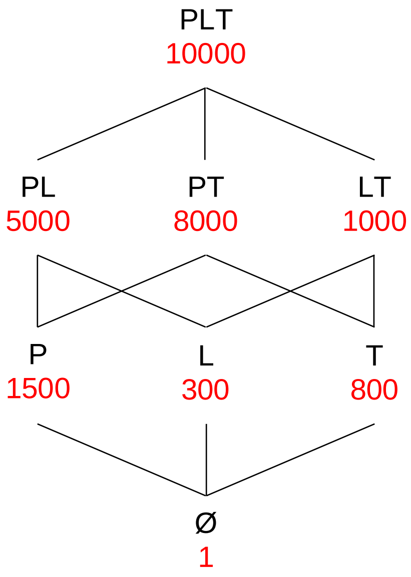
Question 2 [11 marks] Views and materialized views are important techniques used in various data management systems.

(a) [4 marks] Discuss different roles that views play in the following systems. (You should give at least one type of use for each system)

- Relational database systems
- Distributed database systems
- Data warehousing systems
- Federated database systems

(b) [3 marks] Materialized views are pre-computed and stored on the disk. List at least three issues we should consider when materializing views in a data warehousing system.

(c) [4 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?



Question 3 [8 marks] A data warehouse is usually represented by either a star schema or a snowflake schema. Various OLAP operations, e.g., roll-up, drill-down, slice, dice, pivot, etc., can be performed on a data warehouse.

(a) [2 marks] For the following snowflake schema, show the equivalent star schema with dimension tables.

Olympiad Table
Olympiad
City
Organizing Committee
Contact address

Venue Table
Venue
Location
Region

Gender Table
Gender

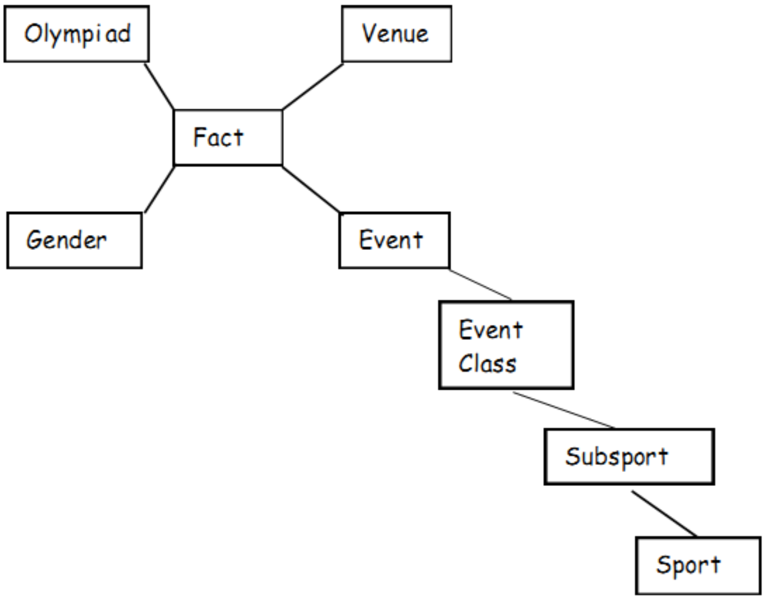
Event Table
Event
Event Class

Event Class Table
Event Class
Subsport

Subsport Table
Subsport
Sport

Sport Table
Sport
Sporting Federation

Fact Table
Olympiad
Venue
Event
Gender
Attendance



(b) [2 marks] What are the advantages and disadvantages of star schema, compared with snowflake schema?

(c) [4 marks] Given the fact table and dimensions in the above snowflake schema, how can we know the attendance of each sport at each venue? (Please explain which OLAP operations are performed on which dimensions in either SQL or plain English)

Question 4. [11 marks] MapReduce is a programming model for processing large-scale datasets with a parallel, distributed algorithm on a cluster of computers. Assume that we have a large census dataset recording information including individual's name, gender, date_of_birth, and country_of_birth. Now we want to use MapReduce to find out the number of people born in each country from the census dataset.

(a) [8 marks] Describe what you need to do in map() and reduce() functions respectively, including the input, output, and the main processing task. (You can use plain English or any programming language).

(b) [1 mark] Explain one way to exchange data between map tasks and reduce tasks.

(c) [2 marks] Some map tasks or reduce tasks may fail due to problems with the node on which they are running. How can such a problem be detected, and how can the MapReduce framework address this problem?

Question 5 [13 marks] Data integration and data quality management are two important pre-processing steps in data warehousing and data mining.

(a) [3 marks] Data quality can be measured from various dimensions. Please explain the meaning of the following data quality dimensions respectively, and give one example of data quality problems for each of these dimensions.

- Accuracy
- Completeness
- Currency

(b) [4 marks] We might encounter various challenges in data integration, especially when we need to link records from different datasets. Please explain the meaning of the following challenges respectively, and give one example for each challenge.

- Schema heterogeneity
- Data type heterogeneity
- Value heterogeneity
- Semantic heterogeneity

(c) [6 marks] Given two strings “Department Engineering” and “Engineering Department”, calculate their similarity using the following string matching techniques. Which technique is relatively more suitable to match these two strings and why?

- Jaccard Coefficient (using 3-gram)
- Edit distance/Levenshtein Metric

Question 6 [6 marks] The big data era has witnessed new challenges in data publishing, data processing and data analysis. Please use techniques you learnt from this course to explain how to address the following issues. (You should list the name of the technique and explain how to address the issue using the technique)

(a) [2 marks] We need to process large-scale data within interactive response time.

(b) [2 marks] Information required for analysis is scattered in multiple data sources.

(c) [2 marks] We need to guarantee privacy in data publishing.

END OF EXAMINATION