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, 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:

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.

For multiple choice questions, please circle only one answer.

Total is 100 marks.

For Examiner Use Only

Question	Mark

Total _____

Question 1. Data Warehouse (19 marks)

Q1.1 (4 marks) What are the main differences between an operational database and a data warehouse?

Q1.2 (4 marks) A data warehouse can often make use of materialized views. Discuss the advantages and disadvantages of building materialized views in data warehouses.

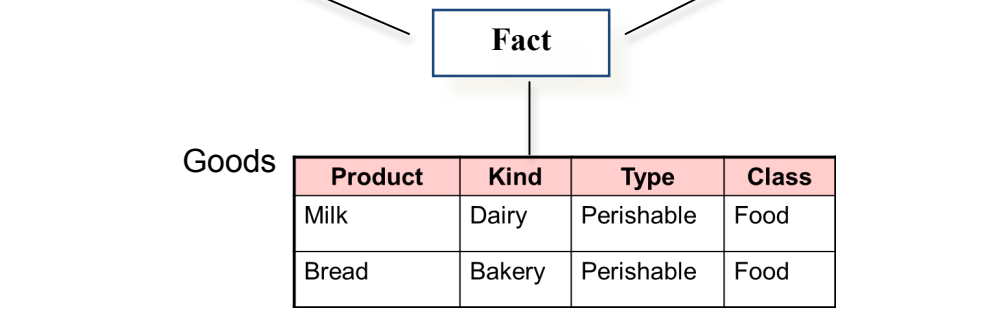
The following is a star schema with a fact table and three dimension tables on *Time*, *Location*, and *Goods*.

Time

Day	Month	Qtr	Year
9.2.04	Feb	1	2004
10.2.04	Feb	1	2004

Location

Store	District	Region
Toowong	North	Brisbane
Sunnybank	South	Brisbane



Goods

Product	Kind	Type	Class
Milk	Dairy	Perishable	Food
Bread	Bakery	Perishable	Food

Q1.3 (3 marks) For the above star schema, show the equivalent snowflake schema with dimension tables.

Assume the population of the fact table is:

	Keys		Facts
Day	**Product**	**Store**	**Sales (AUD)**
9.2.04	Milk	Toowong	20
10.2.04	Milk	Toowong	10
9.2.04	Bread	Toowong	20
10.2.04	Bread	Toowong	30
9.2.04	Milk	Sunnybank	10
10.2.04	Milk	Sunnybank	50
9.2.04	Bread	Sunnybank	10
10.2.04	Bread	Sunnybank	30

Q1.4 (4 marks) Show the result after a rolling up operation on the fact table from *Store* to *Region*, using the *SUM()* aggregate function.

Q1.5 (4 marks) Show the result of pivoting on *Location* dimension and *Time* dimension, using the *SUM()* aggregate function.

Question 2. Data Integration (14 marks)

Q2.1 (2 marks) Provide at least four examples of scenarios in which data integration is needed.

Q2.2 (4 marks) List at least four possible challenges in data integration, and give one example of each challenge.

Q2.3 (8 marks) Health care services in Australia are provided by General Practitioner (GP) clinics and hospitals. All GPs use one database system GPDB and all hospitals use another database system HDB. The schemas of GPDB and HDB are given below:

GPDB: Visit(Medicare#, DoctorID, ClinicID, Date, ConsultationType)

HDB: Admission(pID, InDate, HospitalID, OutDate)

One third-party, the Australian Health Commission (AHC), is responsible to issue a unique ID for each patient. It is a requirement that all patients in GPDB and HDB to register in the AHC patient database.

AHC: Patient(ID, Name, DateOfBirth, HomeAddress, Medicare#)

The GPDB uses a person's Medicare# to identify a patient, and the HDB database uses the ID issued by the AHC for that purpose. For a new patient without an AHC ID, the clinic or hospital needs to apply for an ID for that patient from the AHC.

These three databases are independent and maintained by different bodies in different locations. We want to integrate GPDB and HDB to create the following global view to show all medical visits of a patient:

MedicalVisit(personID,Name,DateOfBirth,VisitDate,VisitType)

Where personID is the ID in the AHC database, and VisitType is either "GP" or "Hospital".

Write an SQL query to generate MedicalVisit.

Question 3. Data Quality (22 marks)

Q3.1 (4 marks) There are many data quality dimensions. Explain the meaning of the following four data quality dimensions, and give one example of data quality problems for each of these four dimensions.

Accuracy:

Representational Consistency:

Currency:

Accessibility:

Q3.2 (4 marks) Data linkage is an operation to identify records referring to the same real-world entity. Why is data linkage so difficult in practice?

Q3.3 (8 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

Q3.4 (6 marks) Compute the edit distance between the two strings “VIECLE” and “VEHICLE” using dynamic programming algorithm. Show the process step by step in the following matrix.

		V	**E**	**H**	**I**	**C**	**L**	**E**
	0	1	2	3	4	5	6	7
V	1							
I	2							
E	3							
C	4							
L	5							
E	6							

Question 4. Multiple Choice Questions (3 marks * 15 = 45 marks)

Q4.1 Which of the following statements is true for warehouse-driven data integration:

A. Data is not duplicated

B. There is a delay in query processing

C. High query performance is achieved

Q4.2 At which granularity level should facts be stored in the multidimensional model?

A. Lowest (finest) granularity

B. Average granularity

C. Highest (coarsest) granularity

Q4.3 Compared to the star schema, the snowflake schema

A. Has de-normalized dimension tables

B. Has a better performance

C. Is harder to use due to many joins

Q4.4 The use of shared dimensions is important for

A. The design of data marts that can be integrated

B. Increasing the query performance

C. Breaking down the development process into small chunks

Q4.5 Pre-aggregation in DW is done to

A. Reduce the space requirements

B. Improve query performance

C. Both A & B

Q4.6 How many pre-aggregates can be computed in an *n*-dimensional data cube?

A. 2^n

B. n^2

C. n

Q4.7 In the greedy algorithm for selecting which views to materialize in a DW, the benefit of a view *v* depends on:

A. Only the views that depend on *v* (i.e., the views that can be derived from *v*)

B. The set of all possible views in the lattice

C. The set of already selected materialized views and the views that depend on *v*

Q4.8 Which of the following statements is NOT true for data integration?

A. Data integration brings data into a physically centralized database

B. A global user can access information from multiple DBs as if they were accessing a single centralized database

C. Local DBs maintain their local autonomy

Q4.9 Which of the following is NOT a necessary step in data integration?

A. Schema mapping

B. Outlier detection

C. Data mapping

Q4.10 Assume *A*<*B* means that system *B* is more strongly integrated than system *A*. Accordingly, which of the following rankings on strength of integration is correct?

A. Distributed DB < Interoperable Systems < Federated DB

B. Federated DB < Distributed DB < Interoperable Systems

C. Distributed DB < Federated DB < Interoperable Systems

Q4.11 For two strings with *m* and *n* characters respectively, the maximum possible edit distance is

A. max(m, n)

B. m+n

C. m*n

Q4.12 For two strings with *m* and *n* characters respectively, the minimum possible edit distance is

A. max(m, n)

B. min(m, n)

C. |m-n|

Q4.13 Which of the following statements about edit distance (ED) between strings *a* and *b* is NOT true?

A. ED(a, b) = 0 if and only if a=b

B. ED(a, b) = ED(b, a)

C. ED(a, c) ≥ ED(a, b) + ED(b, c)

Q4.14 The quality of a relational schema can be measured by

A. Integrity constraints

B. Normal forms

C. Both A and B

Q4.15 For a string of *n* characters, how many *q*-grams it contains?

A. n-q+1

B. n+q-1

C. n*q

END OF EXAMINATION