Preview Test: INFS3200 and INFS7907 Semester One Final Examination 2020

Test Information

Description	INFS3200/7907 Advanced Database Systems Semester One 2020 - Final Examination Exam Duration: 120 minutes + 30 mins additional time Materials permitted include 2 sheets of blank A4 paper for reviewing and planning your approach to the exam, water bottle and Casio fx-82 series or UQ approved and labelled calculator.
Instructions	Answer all questions. There are 50 marks in total. If you experience any technical difficulties during the exam, talk to your online invigilator via the webcam or chat functions. If the technical trouble cannot be resolved, you should ask for an email (or transcript of the chat) documenting any technical advice provided to support your request for a deferred exam.
Timed Test	This test has a time limit of 2 hours and 30 minutes.This test will save and be submitted automatically when the time expires. Warnings appear when **half the time**, **5 minutes**, **1 minute**, and **30 seconds** remain. *[The timer does not appear when previewing this test]*
Multiple Attempts	Not allowed. This Test can only be taken once.
Force Completion	This Test can be saved and resumed at any point until time has expired. The timer will continue to run if you leave the test. Your answers are saved automatically.

QUESTION 1

0 points Save Answer

Please use this space to specify any assumptions you have made in completing the exam and which questions those assumptions relate to. You may also include queries you may have made with respect to a particular question, should you have been able to 'raise your hand' in an examination room.

QUESTION 2

9 points Save Answer

A semijoin is a special type of join operation that can be used in distributed database design (Derived horizontal fragmentation), distributed query processing (replace a join with semijoins), etc.

(a) [**3 marks**] When performing fragmentation on a relation, what properties should the fragmentation meet to ensure the correctness? List their names and meanings.

(b) [**2 marks**] For any join query in a distributed database, does a semijoin always have less or equal join cost than a traditional join? Simply answer "Yes" or "No".

(c) [**4 marks**] Consider two relations R(A, B) and S(X, A, C), where S.A is the foreign key. Assume that the relation R(A, B) is located on site 1 and the relation S(X, A, C) is located on site 2. Consider a join query $R \bowtie_A S$ issued at site 1. Please give a step-by-step query execution plan using semijoin operations to process this query (using either plain English or SQL queries). (**Note:** Use plain English if you cannot find join and semijoin symbols on Blackboard)

QUESTION 3

5 points Save Answer

(a) [**2 marks**] A data warehouse is usually represented by either a star schema or a snowflake schema. What are the advantages and disadvantages of a star schema compared with a snowflake schema?

(b) [**3 marks**] It is not common for data warehousing systems to support update operations. Describe one reason why supporting updates in data warehouses is not a good idea. Briefly justify your answer.

QUESTION 4

12 points Save Answer

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

(a) [**4 marks**] In general, for a 4-dimensional data cube with no hierarchical structure, how many materialized cuboids can be pre-computed? What if one of the dimensions contains three levels (e.g. *location* < *city* < *country* for *location* dimension)?

(b) [**2 marks**] Suppose that a cube contains three dimensions, namely *product (P), location (L) and time (T),* and the cuboid on *{product, location}* is materialized. Among the following group-by queries, which queries can benefit from this materialized cuboid?

{product, location, time}

{product, location}

{product, time}

{location, time}

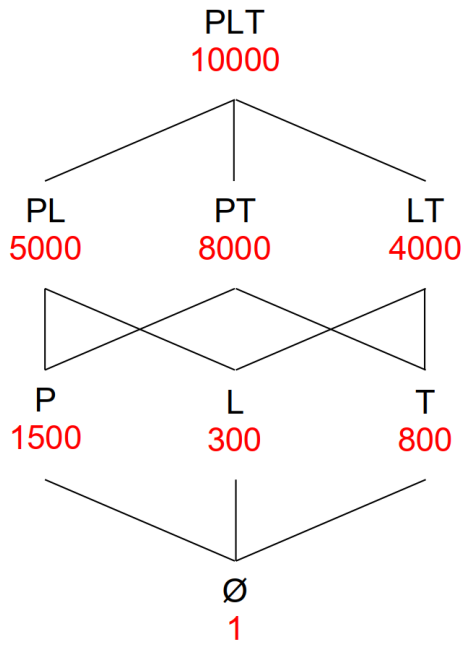
{product}

{location}

{time}

Ø

(c) [**6 marks**] Given the above cube with three dimensions (P, L, 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.



[**INFS3200 ONLY**] Assume that all the queries are issued with the same frequency, what are the first two cuboids that should be materialized in order to minimize the total query cost and why?

[**INFS7907 ONLY**] Assume 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 the total query cost and why?

QUESTION 5

5 points Save Answer

(a) [**2 marks**] List at least two different roles a *database view* plays in different types of systems (one role in each system). Specify the name of each system type and its role in the system.

(b) [**3 marks**] List at least three data heterogeneity challenges when integrating data from different sources. List their names and meanings.

QUESTION 6

7 points Save Answer

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

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

QUESTION 7

6 points Save Answer

Data privacy is a very important issue when publishing data. K-anonymity is a common and simple solution to privacy-preserving data publishing.

(a) [**2 marks**] What is k-anonymity?

(b) **[2 marks**] 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 its difference with K-anonymity.

QUESTION 8

6 points Save Answer

The big data era has witnessed new challenges in data publishing, data processing, and data analysis. Based on the techniques you learnt from this course, choose the best technique for the following application where centralised RDBMS is not applicable. Please list the name of each technique.

(a) [**2 marks**] We need to handle massive flight ticket booking requests concurrently.

(b) [**2 marks**] We need to decide the next move for our company based on the analysis of our historical data.

(b) [**2 marks**] We need to count the appearance of the word “virus” in all books in our library efficiently.