Final Exam: Advanced Database Systems WINTER (COMP90050_2022_WIN)

Started: Aug 2 at 14:00

Quiz Instructions

Total number of questions: 11

Total marks: 50

Authorised Materials:

This is an open book exam – the lecture slides, lecture recordings, textbooks, and study materials can be used. Note that there is a strict time limit for the exam, hence you should be mindful of the time spent using such resources.

While you are undertaking this exam you ***must not***:

- make use of any messaging or communications technology.
- act in any manner that could be regarded as providing assistance to another student who is undertaking this assessment, or will in the future be undertaking this assessment.

The work you submit ***must be based on your own knowledge and skills***, without assistance from any other person.

Instructions to Students:

This exam begins at the time mentioned in 'Available from' in Canvas (lms.unimelb.edu.au). The exam must be completed and submitted by the mentioned 'Due time'. This exam has a strict time limit.

- **Once submitted, you will NOT be able to reopen the quiz to change your answers. This means that you will be able to click the "Submit" button ONLY once.**
- Answers should only contain simple text. You do NOT need to upload any image. DO NOT click anything in the editing toolbar, e.g., changing format, uploading media, adding equation, etc.
- **Questions have unequal marks.** There is no negative marking, and we encourage you to **attempt all questions.**

Academic Integrity Declaration

By commencing and/or submitting this assessment I agree that I have read and understood the **University's policy on academic integrity. (https://academicintegrity.unimelb.edu.au/#online-exams)**

I also agree that:

1. Unless paragraph 2 applies, the work I submit will be original and solely my own work (cheating);

2. I will not seek or receive any assistance from any other person (collusion) except where the work is for a designated collaborative task, in which case the individual contributions will be indicated; and,
3. I will not use any sources without proper acknowledgment or referencing (plagiarism).
4. Where the work I submit is a computer program or code, I will ensure that:
   a. any code I have copied is clearly noted by identifying the source of that code at the start of the program or in a header file or, that comments inline identify the start and end of the copied code; and
   b. any modifications to code sourced from elsewhere will be commented upon to show the nature of the modification.

Question 1

2 pts

Jane has developed an online game which does not have many users at this moment. She is expecting a rapid growth in the number of users in near future, but cannot confirm the exact number or the time of growth. Currently the game is hosted in a server placed in Jane's garage. Jane does not have the space to accommodate more servers in her garage, but needs more computing and storage capacity for the game if a growth happens. Which one of the following solutions is the most suitable (cost, profit, and space effective) choice for this scenario?

- ◯ Establish a cluster for higher computing and storage capacity
- ◯ Buy more machines, place them somewhere else and establish a distributed system
- ◯ If the system is about to overload, refuse service to some users
- ◯ Buy cloud computing and storage services

Question 2

2 pts

Alice has just shared an interesting story in her social network and asked her friend Jane to check. Jane logs in her account but cannot find the story there. As Alice can see her own story from her account every time she checks, she asks Jane to wait for some time. After a minute, Jane checks back and finds the story that Alice shared. What one of the following statements is false for this scenario?

- ◯ That social network is using strong consistency model
- ◯ That social network has a higher preference for availability over consistency

- ◯ That social network is using read-your-write consistency model
- ◯ That social network is using eventual consistency model

Question 3

2 pts

A system uses logged writes to ensure consistency of disk writes. Given a data block where some modification is made, which one of the following options is true?

- ◯ Logged write takes more time (i.e., less efficient) compared to duplex writes.
- ◯ The modification needs to be logged first before the updated data block is written on disk
- ◯ Both the data block and the log about the modification made to that data block, need to be written twice in two different places
- ◯ If the system fails to log about the modification made to that data block for some reason, still the data block can be written to disk.

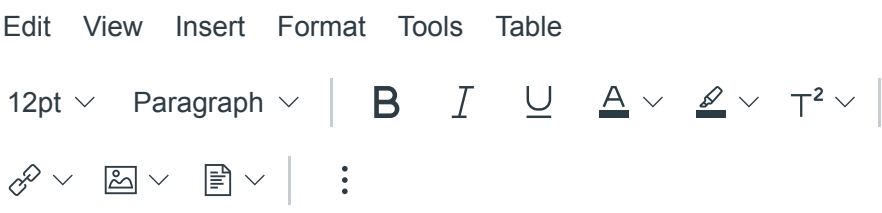
Question 4

5 pts

John needs to buy a new computer. He usually uses his computer to play games, listen to music, watch movies, work on word documents, and browse social networks. He cannot decide which of the following options of hardware will be suitable for him. Can you please explain the advantages and disadvantages of each option for him? Assume both options cost the same.

Option 1: 500 GB hard disk drive, 32 MB cache

Option 2: 250 GB solid state drive, 16 MB cache



p 0 words </>

Question 5 5 pts

A failfast system has 8 devices. Assume 5 out of the 8 devices are unavailable but there are still 2 agreeing devices. Can the system continue to operate in this scenario? Please explain your answer.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 6 5 pts

Let's assume the following information for a bank database with two tables, customer and depositor:

Number of records of *customer*: 1000

Number of blocks of *customer*: 100

Number of records of *depositor*: 500

Number of blocks of *depositor*: 50

A block nested loop join needs to be performed on the customer and depositor tables. Assume that there is enough memory only to hold one block of each table. Which table should be chosen as the outer relation of the join so that the total number of block transfers is minimum? Please explain your answer. You are not required to show exact numeric values of any calculation.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 7

5 pts

For the following data and query, which one of the indexes that we have learnt about in this subject, is a suitable choice? If there is any search key that needs to

be specified for that index, please mention the search key as well. Please explain the rationale of your choice.

The data is the records of one million families, where each family is a row in a table. A unique ID and the total household income per year of each family are stored in the database. The total household income of any family can be between 0 to 2 million, and it is an integer number (i.e., a whole number without any decimal value). The most common queries in this database look like – (i) find the IDs of the families where the total household income is less than 70,000; (ii) find the IDs of the families where the total household income is between 70,000 and 100,000; (iii) find the IDs of the families where the total household income exactly equals to 72,050;

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words

Question 8 6 pts

What are the benefits of using snapshot isolation, compared to two-phase well-formed isolation? Also, give an example of an application scenario where snapshot isolation can be useful over a two-phase well-formed isolation.

Edit View Insert Format Tools Table

12pt Paragraph

12pt Paragraph

p 0 words

Question 9

6 pts

Given a nested transaction where the parent transaction is P, the children transactions of P are T1 and T2, please answer for both scenarios.

(i) Scenario 1: When the value of the variable some_input is 3, which of the following transactions can run concurrently from the beginning till end (that is, all operations and locks are compatible to run concurrently with another one) and which ones need to be delayed? Please give explanation for the delayed transactions.

(ii) Scenario 2: When the value of the variable some_input is 1, which of the following transactions can run concurrently from the beginning till end (that is, all operations and locks are compatible to run concurrently with another one) and which ones need to be delayed? Please give explanation for the delayed transactions.

A compatibility matrix is as follows-

Compatibility Mode of Granular Locks							
Current	None	IS	IX	S	SIX	U	X
Request	+I- (Next mode) + granted / - delayed						
IS	+(IS)	+(IS)	+(IX)	+(S)	+(SIX)	-(U)	-(X)
IX	+(IX)	+(IX)	+(IX)	-(S)	-(SIX)	-(U)	-(X)
S	+(S)	+(S)	-(IX)	+(S)	-(SIX)	-(U)	-(X)
SIX	+(SIX)	+(SIX)	-(IX)	-(S)	-(SIX)	-(U)	-(X)
U	+(U)	+(U)	-(IX)	+(U)	-(SIX)	-(U)	-(X)
X	+(X)	-(IS)	-(IX)	-(S)	-(SIX)	-(U)	-(X)

	T1	T2
P	BEGIN	BEGIN
BEGIN	Lock (IS,A)	Lock (IX,A)
Start both T1 and T2	If(some_input == 3){	If(some_input == 3){
Lock (S,A)	Lock(S,A)	Lock (X,A)
Read A	Read A	Write A
Unlock A	}	}
END	Unlock A	Unlock A
	END	END

Question 10

6 pts

A researcher collected 10 Terabyte of data for her research. She needs to choose a RAID structure to store the data. She needs high reliability, high read throughput, and cheaper option for her work, but write throughput is not a concern for her. She is undecided between the following two options:

Option 1: RAID 1 with 2 disks

Option 2: RAID 4 with 3 disks.

Assuming each disk has 10 Terabyte capacity and the same mean time to failure, can you please explain the advantages and disadvantages of each option for her? If one of these options is clearly a more suitable choice over the other, please explain why. If not, please explain that as well.

Question 11

6 pts

We have a simplified log at the time of a system crash. Assume that there is no log record before the checkpoint. The format of a log record is (LSN, Operation Details).

(00, begin checkpoint)

(05, end checkpoint)

(10, T1 write page1)

(20, T2 write page2)

(30, T1 write page5)

(40, T2 commit)

(45, T2 end)

(50, T3 write page5)

(60, T1 abort)

(70, CLR undo T1 LSN 30)

(80, CLR undo T1 LSN 10)

(90, T1 end)

CRASH

The system recovery consists of three phases: analysis, redo and undo. Please answer each of the following questions.

A. What information will be in the dirty page table after the analysis phase (write as a list of the format (Page id, LSN))?

B. If the pageLSN of Page5 stored in the database is found as **30**, then what will be the order of the LSNs to be redone in the Redo phase? Assume that all the necessary pages are in the dirty page table, all LSNs in the log are greater than or equal to the corresponding page's recLSN, and all LSNs in the log are greater than the corresponding page's pageLSN (except for Page5).

C. What is the order of the LSNs to undo in the Undo phase?

Not saved Submit Quiz