Xi’an Jiaotong-Liverpool University

PAPER CODE	EXAMINER	DEPARTMENT	TEL
CPT201		Computing	

FIRST SEMESTER 2020/2021 Resit EXAMINATIONS

BACHELOR DEGREE – Year 3

DATABASE DEVELOPMENT AND DESIGN

Exam Duration: *2 Hours*

INSTRUCTIONS TO CANDIDATES

1、 Total marks available are 100. This will count for 100% in the final assessment.

2、 Answer ALL FOUR questions.

3、 The number in the column on the right indicates the marks for each section.

4、 The university approved calculator - Casio FS82ES/83ES can be used.

5、 All the answers must be in English.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Question 1. Suppose that a relation called *student* holds 25,000 tuples, which are stored as fixed length and fixed format records. The length of each tuple is 350 bytes. The key attribute, *student_ID*, occupies 10 bytes and another attribute *address* occupies 50 bytes. The records are sequentially ordered by *student_ID* and stored in a number of blocks. Each block has the size of 4,096 bytes (i.e., 4 Kilobytes). Assume that a complete record or an index entry must be stored in one block.

a) How many tuples of the relation *student* can one block hold?

[2/25]

b) How many blocks are needed to store the relation *student*?

[4/25]

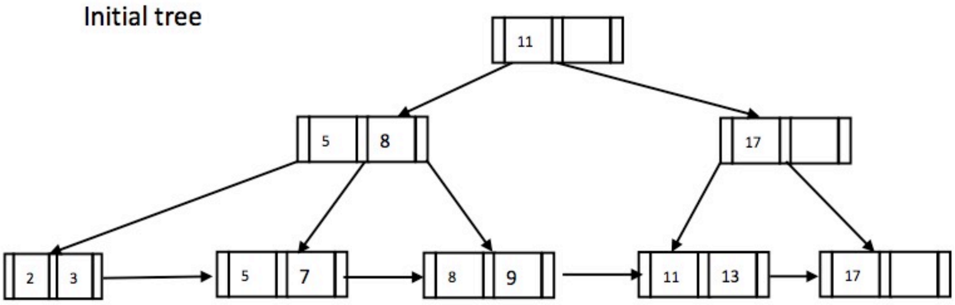
c) Consider creating a primary index on the *student_ID* attribute. Each index entry contains a search key and a 10-byte long pointer to the records. Suppose the primary index is sparse (i.e. one index entry for one block), compute the number of blocks needed to store the index.

[4/25]

d) Briefly explain under what circumstances the two strategies, "merge siblings" and "redistribute pointers" should be used during deletion operations in a B+ tree.

[3/25]

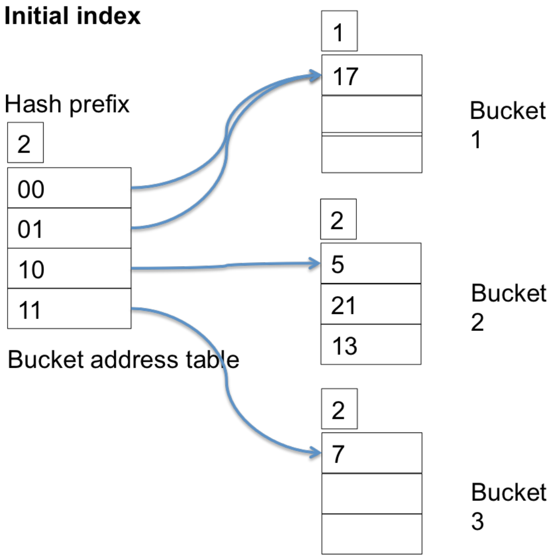
e) Consider the initial B+ tree index shown below. The number of pointers in a node is 3. Draw the B+ tree for each of following operations (in total there should be two trees): (1) insert '10'; (2) delete '17'. (Each of the operations, besides (1), must be performed on the following B+ tree drawn for the previous operation and (1) must be performed on the following B+ tree).



[6/25]

f) Suppose that an extendable hash index is built with the hash function h(x) = x mod 8, and a bucket can hold up to three tuples. The initial hash index is shown below. Draw one hash index for each of the following operations: (1) insert 2, (2) insert 37. (Each of the operations, besides (1), must be performed on the index drawn for the previous operation and (1) must be performed on the following initial index.)

[6/25]


Initial index
Hash prefix
2
00
01
10
11
Bucket address table
1
17
Bucket 1
2
5
21
13
Bucket 2
2
7
Bucket 3

Question 2. Answer the following questions on query evaluation and optimisation. Consider the following two relations.

i) *customer(customer_ID, customer_Name, city, email, account_Number)*

ii) *account(account_Number, branch_Name, balance)*

The *customer_ID* and *account_Number* are the candidate keys for the relations *customer* and *account*, respectively. The *customer.account_Number* is the foreign referencing *account*. Tuples in *account* are sequentially ordered by *account_Number*. The number of tuples in *customer* is 60,000 and the number of blocks is 3,000. The number of tuples in *account* is 70,000 and the number of blocks is 2,000.

[25 marks]

a) Using *customer* as the outer relation, and the block nested-loop join algorithm to evaluate the natural join account customer, how many block transfers would be needed? How many seeks would be needed?

[4/25]

b) Assume that the memory size M is 50 blocks and four blocks are used to buffer the input and output, i.e., $b_b=4$. To sort the *customer* relation based on *account_Number* using external sort merge algorithm, how many block transfers would be needed? How many seeks would be needed?

[8/25]

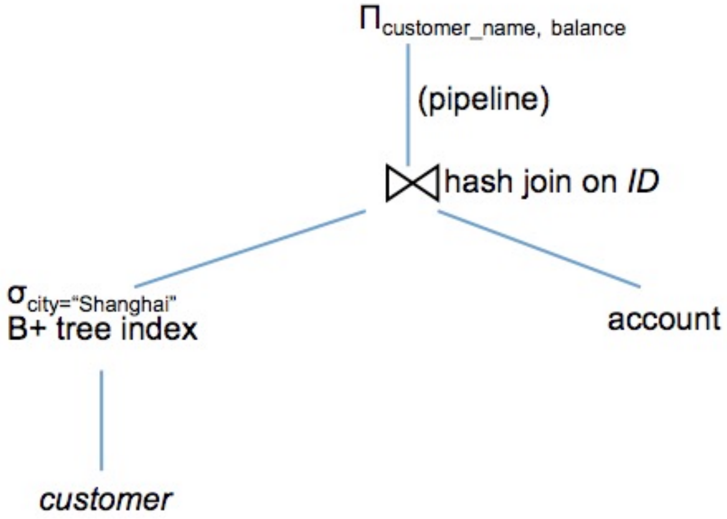
c) Assume that both relations are now physically sorted, to evaluate the natural join *account*$\bowtie$*customer* using the merge join algorithm with $b_b=4$, how many block transfers would be needed? How many seeks would be needed?

[8/25]

d) Additional catalog information about the two relations is given as follows.

- a primary B+-tree index of height 5 on attribute *city* in relation *customer*;
- number of distinct values on attribute *city*, V(*city, customer*) = 1,000;

An optimised evaluation plan for a query is shown below. The hash join algorithm is used to evaluate the join. Assume that the entire build input using the smaller relation can fit in memory. What is the total number of block transfers for the whole evaluation plan? Justify your answer.


Π customer_name, balance
(pipeline)
⋈ hash join on ID
σ city="Shanghai"
B+ tree index
account
customer

[5/25]

Question 3. Answer the following questions related to transaction, concurrency and failure recovery in database systems.

[25 marks]

a) Consider the following schedule.

T1:write(X); T1:write(Y); T2:read(X); T2:write(Y); T2:read(Z); T1:write(Z); T1:commit; T3:read(Y); T2: commit; T3:write(Z); T4:read(Z); T3:commit; T4:abort.

Draw the precedence diagram for the schedule and determine if it is conflict serialisable.

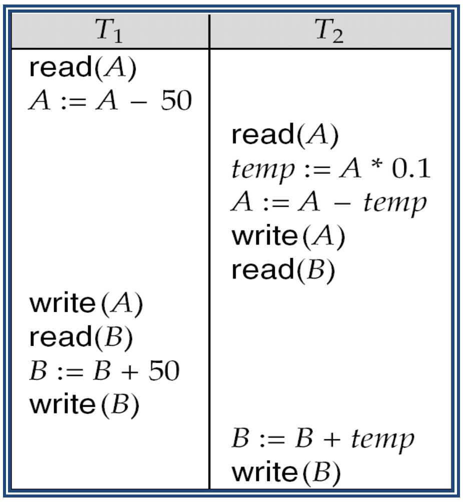
[5/25]

b) Is the schedule in Question 3.a) recoverable? Justify your answer.

[5/25]

c) Consider the following schedule. Can it be transformed to a serial schedule? Justify your answer. Assume that the initial values of A and B are both 1,000. What is the value of A+B after the schedule is executed?

[5/25]



T_1	T_2
read(A)	
$A := A - 50$	
	read(A)
	$temp := A * 0.1$
	$A := A - temp$
	write(A)
	read(B)
write(A)	
read(B)	
$B := B + 50$	
write(B)	
	$B := B + temp$
	write(B)

d) Consider the following schedule which uses the recovery algorithm with redo/undo operations and checkpoints. Database failure happens immediately after time=20. Assume the initial values for X and Y are both 10. Answer the following questions: (1) what transactions would be in the checkpoint L1? (2) What transactions would be in the checkpoint L2? (3) What transactions need to be redone? (4) What transactions need to be undone? (5) What logs would be added as the result of successful recovery?

Time	**T1**	**T2**	**T3**
0	start		
1	read(X)		
2			start
3			read(Z)
4			Z=Z+1

5	X=X+1		
6		start	
7		read(A)	
8		read(B)	
9		B=B+A	
10	---------------------*Checkpoint L1{}*---------------------		
11			write(Z)
12			commit
13		write(B)	
14		commit	
15	read(Y)		
16	Y=Y+X		
17	write(X)		
18	---------------------*Checkpoint L2{}*---------------------		
19	Y=Y+1		
20	***Failure occurs, recovery starts***		

[10/25]

Question 4. Answer the following questions.

[25 Marks]

a) A database contains four transactions as shown in the table below for market basket analysis. Columns represent the transactions and item list.
   (1) Make a tabular and binary representation of the data in order to better view the relationship between items. An item can be treated as a binary variable whose value is "1" if it is present in a transaction and "0" otherwise. And compute the following:
   (2) support (E);
   (3) confidence(A->>EF);
   (4) support(C->>K);
   (5) confidence (K->>L).

[5/25]

Transaction	**Item list**
t1	{L, A, C, F}
t2	{B, C, F}
t3	{F, D, K, A}
t4	{C, A, E, F}

b) The company *Rescue Me!* uses the following three relations in a distributed database at two sites:
   Site 1:
   Adopter (Adopter_ID, Adopter_name, Address, City)
   Adoption (Animal_ID, Adopter_ID, AdoptDate)
   Site 2
   Animals (Animal_ID, Animal_name, PrevOwner, chipNo)
   To answer the query "*Find the name of adopters from the city of Suzhou and their adopted animal names*" one needs to compute the join of the three relations. It is known that only a small number of animals stored at the Site 2 were adopted by people from Suzhou. Assume that a query is made to the Site 1; describe how bloom-join can be used to optimise the query.

[10/25]

c) *Chongqing Tongjunge Drugstores Co. Ltd* - the largest pharmacy store chain in China uses a traditional relational database to store its business data, e.g.

- each patient is identified by an ID number, name, address, and age.
- each doctor is identified by an ID number, name, specialty, and years of experience.
- each pharmaceutical company is identified by name and a phone number.
- each drug has a trade name and a formula. Each drug is sold by a given pharmaceutical company, etc.

As the business goes well and the amount of data keeps growing, it becomes extremely difficult for the company to manage the complex relations among the data items, e.g. a number of similar drugs might have the same provider but with different formula, being prescribed by different doctors for patients.

An IT consultancy company proposes the following solutions to re-design the database system by using:

- semi-structured data store using XML
- linked data (or knowledge graph)

You are asked to make a final choice (only one from the above two), and then describe your design of the data model based on your choice. You need to explain how the data can be modeled and represented with your choice and discuss its advantages. Maximum 100 words.

[10/25]

END OF EXAM PAPER