UNIVERSITY OF COLOMBO, SRI LANKA

UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

DEGREE OF BACHELOR OF INFORMATION TECHNOLOGY (EXTERNAL)

Academic Year 2008/2009 –3rd Year Examination – Semester 6

IT6403 - Database Systems II

Structured Question Paper

13th September, 2009

(TWO HOURS)

To be completed by the candidate

BIT Examination Index No: .....

Important Instructions:

- The duration of the paper is **2 (two) hours**.
- The medium of instruction and questions is English.
- This paper has **4 questions** and **15 pages**.
- **Answer all questions (**25 marks each**)**.
- **Write your answers** in English using the space provided **in this question paper**.
- Do not tear off any part of this answer book.
- Under no circumstances may this book, used or unused, be removed from the Examination Hall by a candidate.
- Note that questions appear on both sides of the paper.
  If a page is not printed, please inform the supervisor immediately.
- **Non-programmable Calculators may be used.**

Questions Answered

Indicate by a cross (×), (e.g. [X]) the numbers of the **four** questions answered.

	Question numbers				
To be completed by the candidate by marking a cross (×).	1	2	3	4	
To be completed by the examiners:					

Index No: ……...

1) (a) Although Basic Timestamp Ordering algorithm causes deadlocks, it ensures that the schedules are both conflict serializable and recoverable. Discuss whether one can agree/disagree with this statement. Justify your answer.

(04 marks)

ANSWER IN THIS BOX

Basic Timestamp Ordering (TO) algorithm does not cause deadlocks

since T waits for T' only if TS(T) > TS(T').

The schedules produced by basic TO are guaranteed to be conflict serializable

since whenever the basic TO algorithm detects two conflicting operations

which occur in the incorrect order, it rejects the later of the two operations

by aborting the transaction that issued it.

The schedules produced by basic TO are not guaranteed to be recoverable and

an additional protocol (i.e. strict TO) must be enforced to ensure that the

schedules are recoverable.

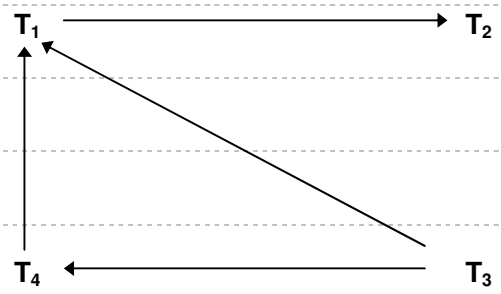
(b)The following schedule consists of four transactions T_1, T_2, T_3 and T_4 where r_i, w_i and c_i means the read, write and commit operations of the transaction Ti respectively.

$r_1(A)$, $w_1(A)$, $r_2(A)$, $r_3(B)$, $w_3(B)$, $w_2(A)$, $r_4(B)$, $w_1(B)$, c_1, c_2, c_4, c_3.

(i) Produce the precedence graph and determine whether the given schedule is conflict serializable. If so give the corresponding serial schedule.

(05 marks)

ANSWER IN THIS BOX



Contd.

The precedence graph does not have a cycle and

hence the schedule is conflict serializable.

The equivalent serial schedule is T_3, T_4, T_1,T_2.

(ii) State giving reasons whether the schedule given above is view serializable.

(02 marks)

ANSWER IN THIS BOX

The given schedule is conflict serializable and hence it is view serializable.

This is due to the reason that all conflict serializable schedules are

view serializable.

(iii) State giving reasons whether the schedule given above is recoverable. If the schedule is not recoverable, give a modified schedule that will be recoverable.

(03 marks)

ANSWER IN THIS BOX

The schedule is not recoverable.

A modified schedule that is recoverable is

$r_1(A)$, $w_1(A)$, $r_2(A)$,

$r_3(B)$, $w_3(B)$, $w_2(A)$, $r_4(B)$,

$w_1(B)$, c_3, c_4, c_1, c_2.

(c) "2PL ensures conflict serializability". Discuss this statement with respect to the schedule given below. Note that r_i, w_i means the read and write operations of the transaction T_i respectively.

$r_1(X), w_2(X), r_3(Y), w_1(Y)$

(04 marks)

ANSWER IN THIS BOX

The given schedule is conflict serializable and it is equivalent to the serial schedule T_3, T_1, T_2.

However, this schedule cannot be produced through 2PL since locks acquired by T_1 cannot be released as there is no deadlock situation for the given schedule.

Although 2PL ensures conflict serializability it does not mean that all conflict serializable schedules can be executed by 2PL.

(d) Consider two Transactions T_1 and T_2 which are received by a system at time 1 and 2, respectively. What will happen if the following schedule is executed according to the timestamp ordering protocol?

$w_1(X), r_2(X), r_2(Y), w_1(Y)$

(04 marks)

ANSWER IN THIS BOX

T_1	T_2
$w_1(X)$	
W-time(X) = 1	
	$r_2(X)$
	R-time(X) = 2

Contd.

$r_2(Y)$

R-time(Y) = 2

$w_1(Y)$

rejected

T1 is rolled back, which causes T_2 to be rolled back too,

because it had read the value of X written by T_1.

(e) Consider the following four recovery concepts.

(i) NO-UNDO / REDO
(ii) UNDO / NO-REDO
(iii) UNDO / REDO
(iv) NO-UNDO / NO-REDO

Name the recovery technique for each of the above recovery concepts.

(03 marks)

ANSWER IN THIS BOX

(i) **Deferred Update**

(ii) **Immediate Update**

(iii) **Immediate Update**

(iv) **Shadow Paging OR**

Deferred Update

2) (a) To create an index on a relation, a choice between hash index and B+-tree index has to be made. For each of the two indexing methods describe when it would be more suitable then the other.

(03 marks)

ANSWER IN THIS BOX

B+-tree index - **If it is likely that ranged queries are going to be performed often,**

then we should use a B+-tree on the index for the relation

since hash indices cannot perform range queries.

Hash index - **If it is more likely that we are only going to perform equality queries,**

for example the case of social security numbers,

then hash indices are the best choice

since they allow for the faster retrieval than B+-trees by 2-3 I/Os per request.

(b) The following are operations which are to be performed on files.

(i) Search for records based on a range of field values and the data would not be modified often.
(ii) Search for records based on a range of field values and the data would be modified often.
(iii) Perform data insertions and deletions often and the order of records does not matter with respect to data retrieval.
(iv) Search for a particular record based on a given field value.

For each of the above operations, which file organization/index is most suited?

(04 marks)

ANSWER IN THIS BOX

(i) **Sorted file organization on the given field**

(ii) **B+ tree index on the given field**

(iii) **Heap file.**

(iv) **Hash indexed on the given field**

(c) Consider the following schema.

```
Employee(Eid, Ename, Salary, Age, Did)
Department(Did, Dname, Floor, Mgrid)
Finance(Did, Budget, Expenses)
```

Assume that each Employee record is 40 bytes long, each Department record is 25 bytes long, and each Finance record is 20 bytes long on average. There are 50,000 tuples in Employee and 5,000 tuples in Department. The file system supports 2,000 byte pages. Each employee is working for one department. Assume uniform distribution of values with respect to any given set of tuples and 20% of employees are earning more than Rs.50,000 while 10% of all employees are over 45 years of age. The company owns five floors in the building and due to uniform distribution, given a set of tuples with Did (department identity), 20% of the tuples of the given set would belong to each floor.

For each of the following queries, which file organization/index choices would you choose to speed them up?

(i) SELECT Ename, Age, Salary FROM Employee;

(03 marks)

ANSWER IN THIS BOX

An unclustered hash index on Ename, Age, Salary attributes of Employee

which is an index only plan

OR

Index will not be created since query requires to access all the Employee records and

the records are accessed using a file scan.

(ii) SELECT Did FROM Department WHERE Floor = 5 AND Mgrid~~Budget~~ < 150~~0000~~;

(03 marks)

ANSWER IN THIS BOX

A clustered dense B+ tree index on <Floor, Mgrid> attributes of Department,

then the records would be ordered on these attributes assisting the given query.

When executing this query, the first record with floor = 5 must be retrieved, and

then the other records with Mgrid < 150 can be found in the order of Mgrid.

This is the best plan for this query and is not an index-only plan.

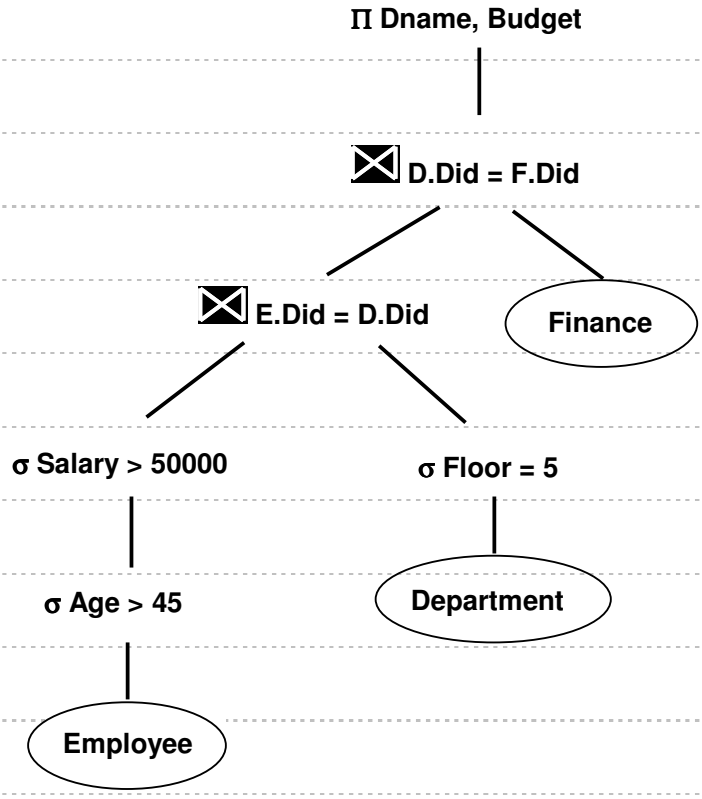
(iii) Consider the following query executed without using any indexes;

```
SELECT D.Dname, F.Budget
FROM Employee E, Department D, Finance F
WHERE E.Did=D.Did AND D.Did=F.Did AND D.Floor = 5
 AND E.Salary > 50000 AND E.Age > 45 ;
```

Draw the optimized query tree for the above query.

(05 marks)

ANSWER IN THIS BOX



(iv) Suggest indices to produce the best plan for the execution of the query in (c)(iii). Explain the query plan illustrating where necessary the number of tuples processed and how the indices would be used in the query plan to minimise the processing cost.

(07 marks)

ANSWER IN THIS BOX

Indices

- **Clustered B+ tree index on age of Employee**
- **Hash/B+ index on Did of Department**
- **Hash/B+ index on Did of Finance**

Contd.

First, retrieve the tuples from Employee with Age > 45

using the B-tree index on Age and based on the estimation 5,000 (10% of 50,000)

such tuples will be found.

The index on Age is more selective than an index on Salary which would have had to

select 10,000 tuples (20% of 50,000).

Out of these retrieved 5,000 tuples, those who earn more than Rs. 50,000

would be selected on the fly. There are 1,000 (20% of 5,000) such tuples.

Pipeline these 1,000 tuples one at a time to Department and using index on Did

it is possible to find at most one matching tuple from the Department tuples

(index nested loop). Consequently, there will be 1,000 tuples.

Out of the retrieved 1,000 tuples, the tuples corresponding to 5th floor

would be selected on the fly and there will be 200 (20% of 1,000) such tuples.

Pipeline the estimated 200 tuples one at a time to Finance and use the index on Did

to retrieve at most one Finance tuple for each of the 200 tuples.

3) (a) (i) In data mining one of the techniques of data analysing and decision making is through Market Basket Analysis? Briefly explain what it is.

(02 marks)

ANSWER IN THIS BOX

Market Basket Analysis is a modelling technique based upon the theory that

if you buy a certain group of items,

you are more (or less) likely to buy another group of items.

(ii) What is an association rule? What is its role in data mining?

(02 marks)

ANSWER IN THIS BOX

An association rule is defined as a statement of the form

{X1,X2,…,Xn} -> {Y1,Y2,...,Yn}, which means that Y1,Y2,...,Yn is present in the

transaction if X1,X2,…,Xn are all in the transaction. It helps to establish relationships

among data sets. For instance the confidence level of a data item against a data set.

(b) A market basket analysis is to be performed to find the relationships between a set of items = {milk, tea, coffee, sugar, juice}. After inspecting eight baskets for these items the following were found.

```
B1 = {milk, tea, sugar}          B2 = {milk, coffee, sugar}
B3 = {milk, coffee, juice}       B4 = {milk, sugar, juice}
B5 = {tea, sugar}                B6 = {tea, sugar, juice}
B7 = {tea, juice}                B8 = {coffee, sugar}
```

(i) Define support in terms of a pair of items A and B. Identify how many of the same pair should be present to achieve a support threshold of 35%.

(03 marks)

ANSWER IN THIS BOX

The support of A=>B is the percentage of the transactions which contains

both A and B.

A pair must appear in 3 of the 8 baskets to have a support of 35%.

(ii) Which pairs of items can be found to meet the support threshold of 35%? Justify your answer.

(03 marks)

ANSWER IN THIS BOX

There are ten possible pairs, each with the following occurrences.

{milk, tea} 1, {milk, coffee} 2, {milk, sugar} 3, {milk, juice} 2,

{tea, coffee} 0, {tea, sugar} 3, {tea, juice} 2,

{coffee, sugar} 2, {coffee, juice} 1,

{sugar, juice} 2.

{milk, sugar} and {tea, sugar} are the pairs which appear 3 times.

(iii) What is the confidence of tea, given milk and sugar (i.e. association rule {milk, sugar} -> tea)?

(02 marks)

ANSWER IN THIS BOX

Milk and sugar appear together in only 3 baskets.

Tea appears in just one of these three.

Thus the confidence of tea given milk and sugar is = 33%.

(c) Consider the following multi dimensional data model of a data warehouse on sales data.

```
Item(ItemId, Name, Category)
Site(SiteId, Town, Country)
Sales(ItemId, SiteId, TimeId, Quantity, Unit-Price)
Time(TimeId, Date, Week, Month, Quarter, Year)
```

(i) What would be the Fact and Dimension tables for the above data warehouse?

(02 marks)

ANSWER IN THIS BOX

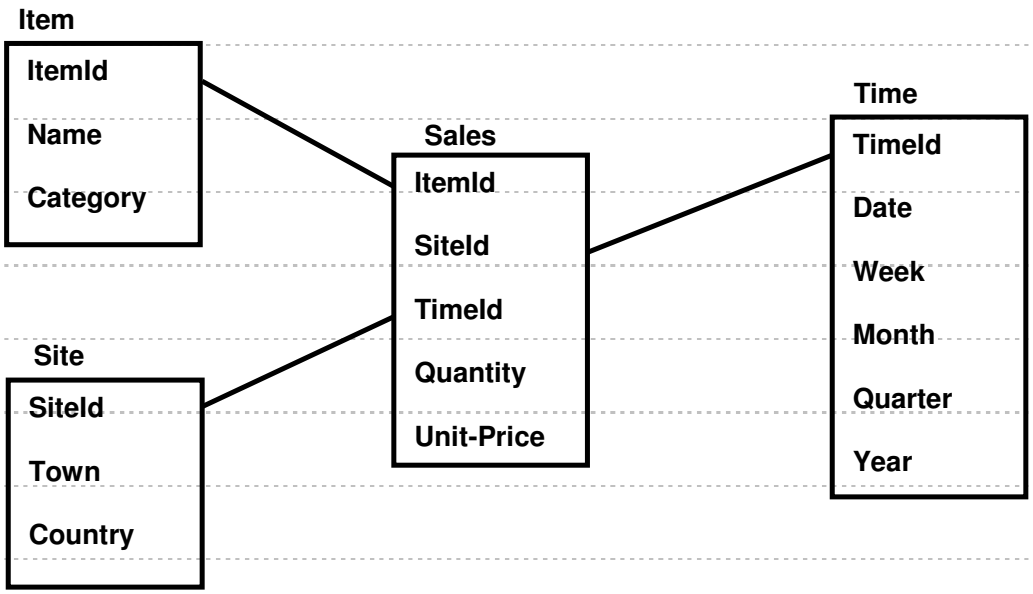
Fact Table: Sales.

Dimension Tables: Item, Site and Time.

(ii) Draw a star schema for the data warehouse in (c)?

(04 marks)

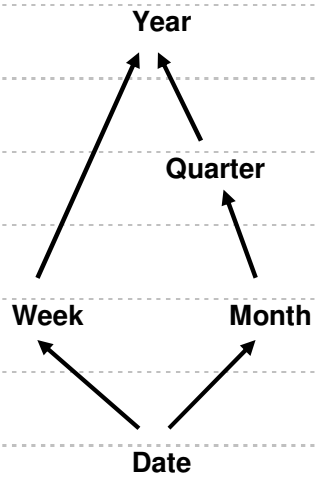
ANSWER IN THIS BOX



(iii) Draw a dimension/concept hierarchy for Time (of the above given data warehouse).

(03 marks)

ANSWER IN THIS BOX



(iv) Write a query to build a cube for the annual Sales for each Town.

(04 marks)

ANSWER IN THIS BOX

```
SELECT Town, Year, SUM(Quantity*Unit-Price)
FROM Sales s, Site si, Time t
WHERE s.SiteId=si.SiteId and s.TimeId=t.TimeId
GROUP BY Town, Year
```

4) (a) Finance department maintains the Employee table to pay salaries of their employees. HR department assigns employees to various projects and maintains Project and Assignment tables. The three tables are defined below.

```
Employee(EmpNo, Name, Title, Salary)
Project(ProjId, Pname, Budget, Location)
Assignment(ProjId, EmpNo, Responsibility, Duration)
```

(i) Most of the projects are based in "Colombo" and hence Location = "Colombo" appears in the majority of queries. Write down the fragments to perform primary fragmentation on Project given the above predicate.

(03 marks)

ANSWER IN THIS BOX

$Proj1 = \sigma_{Location="Colombo"} Project$

$Proj2 = \sigma_{Location<>"Colombo"} Project$

(ii) Write down the fragments if we perform derived horizontal fragmentation on Assignment based on the fragments of Project in (i).

(03 marks)

ANSWER IN THIS BOX

$Asg1 = Assignment \bowtie_{a.ProjId=p.ProjId} Proj1$

$Asg2 = Assignment \bowtie_{a.ProjId=p.ProjId} Proj2$

(iii) Write a query to retrieve all projects with their budgets using the fragmented tables.

(03 marks)

ANSWER IN THIS BOX

```
SELECT Pname, Budget
FROM Proj1
UNION
SELECT Pname, Budget
FROM Proj2
```

(iv) Write a query to identify responsibilities with respect to projects outside "Colombo".

(03 marks)

ANSWER IN THIS BOX

```
SELECT DISTINCT Responsibility  FROM Asg2

OR

SELECT DISTINCT Responsibility FROM Proj2 p, Assignment a
WHERE p.ProjId=a.ProjId
```

(v) Assume that there are two databases called HR and Finance where HR manages Projects and Assignments while Finance manages Employee data.

(I) If these two databases are to be managed as **distributed databases** write a query or query plan to retrieve name with salary of Employees working in projects located in "Colombo".

(04 marks)

ANSWER IN THIS BOX

```
SELECT Name, Salary  FROM Finance.Employee, HR.Asg1
WHERE Finance.Employee.EmpNo=HR.Asg1.EmpNo
```

Note: may use Assignment and Proj1 instead of Asg1 and join them as used in (ii)

(II) If these two databases are to be managed as **multi databases** write a query or query plan to retrieve name with salary of Employees working in projects located in "Colombo".

(04 marks)

ANSWER IN THIS BOX

SELECT Name, Salary FROM Finance.Employee

WHERE Finance.Employee.EmpNo IN

(SELECT HR.Asg1.EmpNo FROM HR.Asg1)

Data from HR Asg1 would be exported and then joined with the Employee table.

(b) What is Embedded SQL?

(02 marks)

ANSWER IN THIS BOX

Embedded SQL is a method of combining the computing power of a

programming language and the database manipulation capabilities of SQL.

(c) Describe Object Definition Language (ODL) and Object Query Language (OQL) used in Object-Oriented Database Management Systems. For each of these two object languages, what is the corresponding language provided by relational database management system called?

(03 marks)

ANSWER IN THIS BOX

Object Definition Language (ODL) is used for defining interfaces to object types.

This is similar to the Data Definition Language (DDL) of Relational DBMS.

OQL is used for the retrieval and manipulation of objects

This is similar to the Data Manipulation Language (DML) of Relational DBMS.

****