UNIVERSITY OF COLOMBO, SRI LANKA

UCSC UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING

DEGREE OF BACHELOR OF INFORMATION TECHNOLOGY
Academic Year 2009/2010 –3rd Year Examination – Semester 6

IT6403 - Database Systems II

Structured Question Paper

1st August, 2010
(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 **14 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 questions answered.

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

1) (a) State the strict two-phase locking protocol.

(03 marks)

ANSWER IN THIS BOX

This is a variant of the two-phase locking protocol which adds the restriction that the shrinking phase does not happen until after the transaction is committed or aborted. As such, based on strict two-phase locking protocol, a transaction must hold all the required locks before executing and does not release any lock until the transaction has completely finished.

(b) Consider the schedule given below.

t_i	T1	T2	T3
1			Read(A)
2			Write(A)
3	Read(B)		
4		Read(C)	
5		Write(C)	
6	Write(B)		
7	Read(C)		
8			Write(B)
9	Commit		
10		Commit	
11			Commit

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

(04 marks)

ANSWER IN THIS BOX

T2 ⟶ T1 ⟶ T3

Schedule is conflict serializable.

Corresponding serial schedule is T2 T1 T3.

(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 interleaving sequence of the above schedule from t_1 to t_9 could be considered by a database.

(03 marks)

ANSWER IN THIS BOX

No, because it is not recoverable.

The reason is that if T1 reads a value of C updated by T2,

T1 commits before T2.

(c) Consider the following schedule where r_i, w_i and c_i mean the read, write and commit operations respectively of the transaction T_i.

$$r_2(x),\ w_3(x),\ c_3\ ,w_1(y),\ c_1,\ r_2(y),\ w_2(y),\ c_2;$$

Assume that T_1, T_2, T_3 have timestamps 1, 2, 3 respectively. Explain whether the above schedule could be executed using timestamp ordering.

(03 marks)

ANSWER IN THIS BOX

This schedule can be executed using timestamp ordering.

When w(x) is issued by T3, r(x) = 2 – no conflict

When r(y) is issued by T2, w(y) = 1 – no conflict

The corresponding serial schedule is T1,T2,T3.

(d) Explain whether the schedule given in (c) above could be executed with 2PL.

(03 marks)

ANSWER IN THIS BOX

No.

T2 must release read lock on x for T3 to obtain write access.

However, this is not possible before T2 obtains locks on y and

hence the violation of two-phase locking protocol.

(e) *"All conflict serializable schedules can be executed by 2PL"*. Justify through an example whether you agree/disagree with this statement.

(03 marks)

ANSWER IN THIS BOX

Although 2PL ensures conflict serializability, it does not mean that all conflict

serializable schedules can be executed by 2PL.

For example, the schedule given in (c) above is conflict serializable

but is not executable by 2PL.

(f) Name one recovery technique for each of the policies given below

(i) No steal no force
(ii) No steal force
(iii) Steal No force
(iv) Steal force

(04 marks)

ANSWER IN THIS BOX

(i) No steal no force - Deferred Update

(ii) No steal force - Shadow Paging

(iii) Steal No force - Immediate Update (Undo/Redo)

(iv) Steal force - Immediate Update (Undo/No Redo)

2) (a) What is an index-only plan with respect to query evaluation and what are the advantages of using index-only plans?

(03 marks)

ANSWER IN THIS BOX

An index-only plan is a query evaluation plan which requires to access only the indexes for the data records and not the data records themselves, in order to answer the query.

As such, the index only plans are much faster than regular plans since it does not require reading of the data records.

If a certain query is executed repeatedly which only require accessing one field (for example the average value of a field), it would be an advantage to create a search key on this field to use an index-only plan.

(b) Consider the following schema for a portion of a simple company database and the SQL statement.

```
Employee(Eid, Ename, Address, Salary, Designation, Deptid)
Department(Deptid, Dname, Budget, Status)

SELECT Designation, Avg(salary)
FROM Employee
GROUP BY Designation
```

Suggest an index only plan for the above query identifying the index type, index attributes and relations involved.

(03 marks)

ANSWER IN THIS BOX

Create a dense B+ tree index on <Designation, Salary>

of the Employee relation.

(c) Assume that the following queries are executed on the company database given above. Suggest the indices that should be created to speed up those query executions. You should specify the corresponding attributes that the index is built on for each index type.

(i) List the *Eid* and *Address* of employees with a user-specified employee name.

(03 marks)

ANSWER IN THIS BOX

Create B+ tree/Hash index on Ename of the Employee relation

(ii) Select the *Eid*, *Ename*, and *Address* of employees who work in the department with a user-specified department name.

(03 marks)

ANSWER IN THIS BOX

create B+Tree / Hash index on Deptid of the Employee relation and

another B+ Tree/Hash index on <Dname, Deptid> in the Department relation.

(iii) List the overall maximum *Salary* for employees.

(03 marks)

ANSWER IN THIS BOX

A B+ tree index on Salary for the Employee relation will help to find the

maximum salary. Since aggregate operation involved is MAX, the proposed index will

enable to traverse down to the rightmost leaf page in the B+ tree index and thus is

better than a hash index.

(d) Consider the same company database.

```
Employee(Eid, Ename, Address, Salary, Designation, Deptid)
Department(Deptid, Dname, Budget, Status_Report)
```

Assume that each Employee record is 100 bytes long and each Department record is 200 bytes long on average. There are 20,000 tuples in Employee and 50 tuples in Department. The file system supports 4000 byte pages. Only 5% of the employees are managers. About 80% of the Departments have a budget greater than Rs. 20,000.

The following questions are based on the information given above. The cost is considered based on *the number of I/O pages*.

(i) Consider the following query:

```
SELECT *
FROM Employee
WHERE Salary > 55000 ;
```

Assume that there is an unclustered B+ index on *Salary*. Let the number of qualifying tuples be *N*. For what values of *N* is a sequential scan cheaper than using the index?

(04 marks)

ANSWER IN THIS BOX

The Employee relation occupies 500 pages. For an unclustered index, the retrieval of N tuples requires N I/O pages. If more than 500 tuples match, the cost of fetching Employee tuples would exceed the cost of sequential scan.

(ii) Consider the following query:

```
SELECT E.Eid, E.Ename, D.Deptid, D.Dname
FROM Employee E, Department  D
WHERE E.Designation = 'Manager' AND D.Budget >20000 AND
E.Deptid=D.Deptid
```

Suggest the indices that should be created to obtain the lowest estimated cost for this query.

(02 marks)

ANSWER IN THIS BOX

A clustered B+ index on E.Designation

A clustered B+ index on D.Budget

(iii) Estimate approximately the corresponding cost based on the best optimized query plan.

(04 marks)

ANSWER IN THIS BOX

With clustered index on Designation of Employee

We get 20,000 * (1/20) {reduction factor} = 1000 tuples (25 I/O)

Write Temp T1 (25 pages) for merge sort

With clustered index on Budget of Department

We get 50 * (80/100) {reduction factor} = 40 tuples (2 I/O)

Write Temp T2 (2 page) for merge sort

Apply merge sort to join T1 and T2.

Merge Sort Cost : T1($2*25*\log_2 25$) + T2($2*2*\log_2 2$) = MS

Total cost = 25+2+MS = 27+MS

3) (a) (i) Knowledge discovery and data mining process is considered to have four stages. Name and briefly explain what each stage is supposed to do.

(04 marks)

ANSWER IN THIS BOX

Data selection - The target subset of data and the attributes of interest are identified by examining the entire raw dataset.

Data cleaning - Noise and outliers are removed, field values are transformed to common units and some new fields are created by combining existing fields to facilitate analysis.

Continued...

Data mining - Data mining algorithms are applied to extract interesting patterns.

Evaluation - The patterns are presented to end-users in an understandable form,

i.e. through visualisation.

(b) Consider the market basket analysis given below to answer the questions (i) to (iii) below. Market basket analysis is to be performed to find the relationships between a set of items = {milk, bread, butter, sugar, tea}. After inspecting four baskets for these items the following were found.

```
B1 = {milk, bread, sugar}        B2 = {milk, butter, sugar}
B3 = {milk, butter, tea}         B4 = {milk, sugar, tea}
```

(i) What is the support of the item set {bread, butter}?

(02 marks)

ANSWER IN THIS BOX

Bread and butter are not contained in any of the transactions.

Hence the support is 0%.

(ii) What is the confidence of having {sugar} in an item set of {milk, tea}?

(02 marks)

ANSWER IN THIS BOX

Confidence = Support of {milk, tea, sugar} / Support of {milk, tea}

= (1 / 4) / (2 / 4) = ½ = 50%.

(iii) Justify the statement "Support of the item set {milk, butter} is similar to the support of the item set {milk, tea}".

(02 marks)

ANSWER IN THIS BOX

Both {milk, butter} and {milk, tea} appear together in exactly 2 baskets.

Hence each of them has a support of 50%.

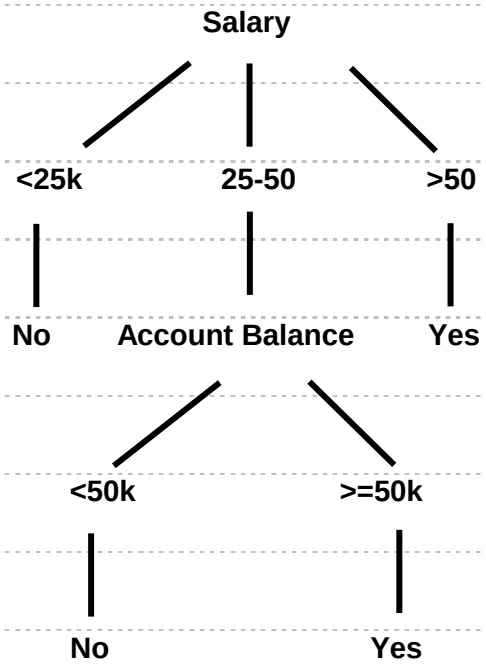
(c) The sample data set is depicting the unworthiness of credit cards with respect to a banking application. It is necessary to build a decision tree based on the given data set.

CID	Age	Salary	Acct Balance	Credit worthy
1	25….55	> 50k	< 50k	Yes
2	25….55	> 50k	>=50k	Yes
3	< 25	25k….50k	< 50k	No
4	< 25	< 25k	>=50k	No
5	> 55	< 25k	>=50k	No
6	> 55	25k….50k	>=50k	Yes
7	< 25	> 50k	< 50k	Yes
8	25….55	25k….50k	< 50k	No

Given that Salary is the highest gain, draw the most appropriate decision tree to partition the above data set to determine the credit worthiness of customers?

(06 marks)

ANSWER IN THIS BOX



(d) Consider the following schema/data model of a data warehouse on sales data.

```
Product(Pid, Pname, Category, Price, Supplier)
Location(Locid, Street, City, Province, Country)
Sale(Pid, Timeid, Locid, Quantity, Price)
Time(Timeid, Date, Week, Month, Quarter, Year)
```

Using the above schema, describe the following data warehouse concepts.

(i) What is/are the dimension(s) of the above data model?

(02 marks)

ANSWER IN THIS BOX

Product, location, Time

(ii) What is/are the fact table(s) of the above data model?

(02 marks)

ANSWER IN THIS BOX

Sale

(iii) If a hierarchy of aggregation levels are used for location, identify the attributes of the above data model that would meet the requirement.

(02 marks)

ANSWER IN THIS BOX

City, Province, Country

(iv) It is required to retrieve the annual revenue for each city. Write an SQL statement for this task.

(03 marks)

ANSWER IN THIS BOX

```
SELECT t.Year, l.City, SUM(s.Quantity*s.Price)
FROM Sale s, Time t, Location l
WHERE s.Timeid=t.Timeid and t.Locid=l.Locid
GROUP BY t.Year, l.City
```

4) (a) Name the fragmentation types and list the advantages of fragmenting a relation based on these types.

(05 marks)

ANSWER IN THIS BOX

Horizontal fragmentation:

- **allows parallel processing on fragments of a relation.**
- **allows a relation to be split, so that most frequently accessed tuples are stored together.**

Vertical fragmentation:

- **allows tuples to be split, so that most frequently accessed attributes are stored together.**

Continued...

- **tuple-id attribute allows efficient joining of vertical fragments allowing parallel processing on a relation.**

Mixed fragmentation:

- **Fragments may be successively fragmented horizontally and vertically to an arbitrary depth to benefit from both schemes.**

(b) What are advantages of replicating a relation in a distributed environment?

(03 marks)

ANSWER IN THIS BOX

Availability:

Failure of site containing the relation does not result in unavailability of data.

Parallelism:

Queries on the relation may be processed by several nodes in parallel.

Reduced data transfer:

The relation may be available locally due to replication.

(c) What are the advantages of location transparency?

(03 marks)

ANSWER IN THIS BOX

Access to remote data is simple because database users do not need to know the physical location of database objects.

Administrators can move database objects with no impact on end-users or existing database applications.

(d) Consider the following two specifications for Address and Employee respectively.

```
CREATE TYPE Addr AS (Street Varchar (45), City Varchar (25),
  House_no Varchar (04));

CREATE TABLE Emp AS (Name Varchar (35), Address Addr, Age
  Integer);
```

(i) What is `Addr` with respect to the above context?

(03 marks)

ANSWER IN THIS BOX

Addr is a user defined type which could be used as an attribute.

Addr could also be defined as a row type of a table.

(ii) Give the SQL syntax to insert an Employee instance.

(04 marks)

ANSWER IN THIS BOX

INSERT INTO Emp ('Perera', addr('Reid Avenue', 'Colombo', '35'), 25);

(e) What is ODMG?

(03 marks)

ANSWER IN THIS BOX

ODMG - Object Data Management Group is a specification for object database and object-relational mapping products.

(f) Discuss how a Relational database is different from an Object-Relational System.

(04 marks)

ANSWER IN THIS BOX

Object-Relational system is built on top of the relational model.

It has an object-oriented data model to support objects, classes and inheritance that are directly supported in the database schemas and in the query language.

Such features are not part of a typical relational model.

****