Sri Lanka Institute of Information Technology

B.Sc. Special Honors Degree

in

Information Technology

Final Examination

Year 3, Semester 2 (2010)

Database Management Systems III (311)

Duration: 3 Hour

9.00 a.m. -- 12.00 noon

16th October 2010.

Instruction to Candidates:

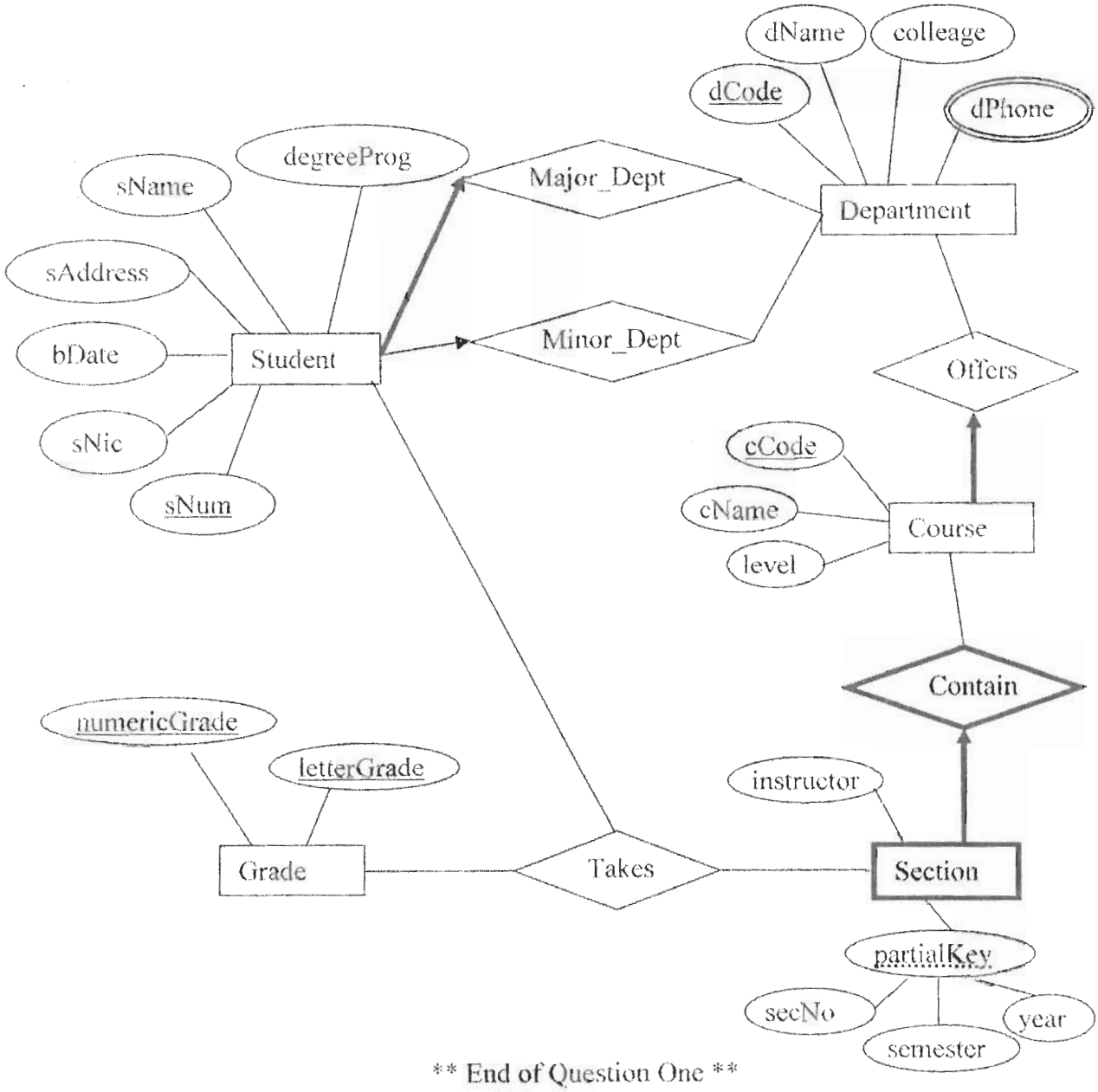
- This paper contains 5 questions. Answer All Questions.
- Provide answers in the booklets given.
- Total Marks for the paper is 100.
- Mark for each question is mentioned in the paper.
- This paper contains 9 pages including Cover Page.

Question 1

Time: 30 minutes
Marks: 20 marks

The Entity-Relationship model given below represents the database requirements of a University database. Map this ER model to an Oracle Object Relational schema. Show the steps in your mapping. Indicate all REF types, VARRAY types, Nested table types, Primary keys, Scoped REFs and any other constraints.

The built-in data types need not be specified. Also, it is not necessary to write CREATE TYPE or CREATE TABLE statements. **(20 marks)**



** End of Question One **

Question 2

Time: 35 minutes
Marks: 20 marks

Consider the following object relational database schema for a University database:

Object types:

dept_type (dno: char(6), dname: varchar(15), campus: varchar(12))
student_type (sid: char(8), sname: varchar(15), dept: ref dept_type, course: varchar(12)).

Tables:

dept of dept_type(dno primary key)
student of student_type(sid primary key)

Dept table stores information about all departments. Dept table contains dept_type objects with attributes department number, department name and campus. The students table stores information about all undergraduate students and graduate students. Student table contain student_type objects with attributes student id, student name, department registered and the course followed.

Assume that these types as well as two tables named dept of dept_type and student of student_type have been created. The attribute dept of the student table references the dept table.

(a) Assume that you are supposed to create two sub types called ***undergrad_type*** and ***grad_type*** under the ***student_type*** to suit the following description.

- The undergrad_type is a subtype of student_type with attributes, (credit_pts: integer, gpa: real) representing the credit points completed and the grade point average to date.
- The subtype grad_type of student_type has attributes, (project: varchar(15), startdate: date, enddate: date) representing the project name, the start date and the end date of the project.

Write Oracle OBJECT SQL statements to create ***undergrad_type*** and ***grad_type.***

(3 marks)

(b) Write Oracle OBJECT SQL statements to answer the following queries (using columns of REF type instead of joins to link tables):

(i) Insert a ***grad_type*** student to student table with the following values. **(2 marks)**

(sid :'2/M1/123', sname:'Sahan', dept:'IT', course:'M.Sc.', project:'Data Mining', startdate:'02-JUN-2010', enddate:'02-JUN-2011')

(ii) Get the student name (sname), course and department name (dname) of graduate students (of grad_type) who started their project before 1 July 2009.

(4 marks)

(iii) Find the names (dname) of departments that have 1000 or more undergraduate students (of undergrad_type). **(4 marks)**

(c) Write a member method called ***rem_time()*** to calculate the remaining time to be spent (in months) by a graduate student on the project. If the student has already finished the project then this method should return '0' otherwise the method must return the number of remaining months (a float value). **(5 marks)**

Hints :

- Function ***months_between(enddate, startdate)*** returns the number of months (a float value) between the given two dates.
- ***SYSDATE*** gives the current system date.

(d) Using the method defined above, write Oracle SQL statement to find the students who has not completed projects yet. Display the name of the student, the name of the project and the time remains. **(2 marks)**

** End of Question Two **

Question 3

Time: 38 minutes
Marks: 20 marks

(a) Define the term *most selective access path* for a given query. **(2 marks)**

(b) If a B+ tree index matches the selection condition, how does clustering affect the cost? Discuss this in terms of the selectivity of the condition. **(2 marks)**

(c) Calculate the cost of sorting Employee relation using External Merge Sort. Employee relation has 1000 tuples with 10 tuples per page. There are 5 buffer pages available. **(3 marks)**

(d) Consider the following BCNF relation schemas

Player (pid, pname, age, phone)
Play (pid, gname, score)
Game (gname, gdate, glocation)

Assume that Player relation consists of 300 pages with 100 tuples per page, Play relation contains 900 pages with 100 tuples per page, and Game relation contains 300 pages with 100 tuples per page. Assume equal-size fields in each relation and uniform distribution of Players for Games in Play relation.

Consider the following query:

Select pr.pid, pr.pname, sum(pl.score)
From Player pr, Play pl
Where pr.pid = pl.pid AND pl.score >25
Group by pr.pid, pr.pname

Assume that 3% of the tuples in Play relation satisfy the selection condition of *score >25* and 50 buffer pages are available.

You are given that the following indexes exits. Assume that a hash index takes 1.2 IOs and a B+ tree index takes 3 IOs to find the rid of a tuple when the key is given.

- Index 1: Unclustered B+ tree index on Play<score, pid>
- Index 2: Unclustered hash index on Player<pid>
- Index 3: Unclustered B+ tree index on Player<pid, pname>

Consider Index Nested Loop Join and Sort Merge Join in choosing an efficient plan for the above query. You may use only the given indexes in considering the best plan.

Note: For simplicity, you may ignore the cost for storing rids in data entries of indexes in your calculations.

(i) Describe two plans based on each of the above join methods. **(6 marks)**

(ii) Show estimations of cost (in disk I/Os) of both plans **(6 marks)**

(iii) Which is the best plan? **(1 marks)**

** End of Question Three **

Question 4
Time: 40 minutes
Marks: 20 marks

(a.) When should we create clustered indexes? **(2 marks)**

(b.) Consider an unclustered hash index on the age field of an employee table. There are 5000 employees who are uniformly distributed in the age range of 21-70. The data records are stored on 500 pages in the order of employee number. If you need to retrieve all employee records where age>30, is it better to use the index? Explain in detail. **(3 marks)**

(c.) Consider the following BCNF relation schema for a portion of a university database

Prof(eid, pname, office, age, sex, speciality, dept_did)
Dept(did, dname, budget, num_majors, chair_eid)

Suppose you know that the following queries are the four most common queries in the workload for this university and that all four are roughly equivalent in frequency and importance.

- List the names, ages and offices of professors of a user-specified sex (male or female) who have a user-specified research specialty (e.g. *recursive query processing*). Assume that the university has a diverse set of faculty members, making it very uncommon for more than few professors to have the same research specialty.
- List department name and chairperson for the departments with a user-specified number of majors. Assume that a considerable number of departments offer the same number of majors.
- List the lowest budget for a department in the university.
- List all information about professors who are chairpersons.

These queries occur much more frequently than updates, so you should build whatever indexes you need to speed up these queries. However, you should not build any unnecessary indexes.

Given this information, design a set of indexes for the university database to give good performance for the expected workload. Decide the indexes (B+ or Hash) to be created and the attributes to be indexed. Mention whether these indexes are clustered or un-clustered.

Mention the reasons to choose each index. (No need to calculate the IO cost) **(5 marks)**

(d.) Mention two differences between Middleware System Architecture and Collaborating Server System Architecture used to build Distributed Databases. **(2 marks)**

(e.) Describe two different ways to update distributed data (replicated data). **(2 marks)**

(f.) Consider the Play and Game relations described below.

Play (pid: integer, gName: varchar, score: integer)
Game (gName: varchar, gLocation: varchar, gDate: date)

They are stored in a distributed DBMS with all Play tuples stored at Perth and all Game tuples stored at Sydney.

Each relation contains 20-byte tuples, and the size of *gName* fields is equal to 5 bytes. The Play relation contains 100,000 pages, the Game relation contains 5,000 pages, and each processor has 200 buffer pages of 4,000 bytes each. The cost of one page I/O is t_d, and the cost of shipping one page is t_s; tuples are shipped in units of one page by waiting for a page to be filled before sending a message from processor i to processor j. There are no indexes, and all joins that are local to a processor are carried out using a sort-merge join

Consider the query:

```
SELECT    *
FROM      Play p , Game g
WHERE     p.gName = g.gName AND
          p.score = 10 ;
```

The query is posed at Colombo, and you are told that only 1 percent of Play tuples satisfy the given condition.

(i) Calculate the result size of the query. **(2 marks)**

(ii) Calculate the cost of computing the query at Perth by shipping Game relation to Perth and then shipping the result to Colombo. **(2 marks)**

(iii) Calculate the cost of computing the query at Perth using Semijoin; then shipping the result to Delhi. **(2 marks)**

** End of Question Four **

Question 5

Time: 37 minutes
Marks: 20 marks

(a.) Mention two motivation factors to built Data Warehouses. **(1 mark)**

(b.) Explain the following terms. **(2 marks)**

(i) ROLAP

(ii) MOLAP

(iii) Roll up

(iv) Drill down

(c.) Briefly explain the concept of Inverted List Indexes and demonstrate how it can be used to find the tuples which match two equality conditions using an example. **(3 marks)**

(d.) Consider the following star schema of a data warehouse:

Suppliers (sno, sname, rating)
Parts (pno, pname, weight)
Projects (jno, jname, city, state, sdate)
Sales (sno, pno, jno, date, quantity, price)

The primary keys are underlined. In the Sales table, sno references Suppliers, pno references Parts, and jno references Projects. The Projects table contains the city and state where each project is located.

Write suitable SQL statements for the following:

(i) Create a materialized view that contains the sales value for each combination of sno, pno, jno and date along with the city and state of each project. (Sales value = quantity*price). **(2 marks)**

(ii) Using the materialized view of (i), get the total sales value for each pno and state. **(1 mark)**

(iii) Modify the SQL statement of (ii) to drill down along the projects dimension to the city level. **(1 mark)**

(e.) Consider the transaction table given below as a database of items bought by various customers from a retail store.

Transaction ID	Items Bought
101	A, B, E
102	A, C, D, E
103	A, C. D
104	B, C. E
105	A, B, C, D

(i) Give all the frequent item sets in this database if the minimum support required is 50%. Show the steps in arriving at your result. **(2 marks)**

(ii) Based on a minimum support of 40% and a minimum confidence of 80%, what are the valid association rules of this database? Give the actual confidence of your rules. **(2 marks)**

(f.) Describe two differences between Classification and Clustering **(2 marks)**

(g.) The following table contains information about a set of customers who visited a computer shop. The last column state whether they purchased a new computer or simply left the shop.

Training Set

Name	Age	Income	Education	Gender	Purchased a new computer
Amy	62	Medium	High school	F	No
Al	53	Low	First Degree	M	Yes
Betty	47	Medium	First Degree	F	No
Bob	32	Medium	Doctorate	M	Yes
Carla	21	High	High school	F	Yes

Test Set

Name	Age	Income	Education	Gender	Purchased a new computer
Nevil	33	Low	High school	M	No
Soh	21	Medium	First Degree	M	Yes
Raj	22	Medium	Masters	M	No
Nethali	33	Medium	High school	F	No
Mary	66	High	Doctorate	F	Yes

(i) Use the first data set as a Training Set to build a model to classify customers as possible buyers or not. Draw a possible decision tree to classify actual data. **(3 marks)**

(ii) Use the second data set as a Test Set and calculate the accuracy of the decision tree you drew in the previous step. **(1 mark)**

** End of Question Five **

END OF QUESTION PAPER