University of Ottawa
CSI 3130 and 3510 – Final Examination
Professor(s): Herna L. Viktor and Iluju Kiringa

2007
14:h00-17h00
Duration: 3 hrs

Closed book; no aid allowed, except one double-sided letter-size “cheat sheet”. Answer all questions in **ink**. **Good luck!**

Family name: ______________________________

First name: ______________________________

Student number: __________________________

There are 8 questions and a total of 100 points.

This exam must contain 9 pages,
including this cover page.

1 – Evaluating Relational Operators	/ 14
2 – External Sorting	/ 10
3 – Query Optimization	/ 10
4 – Concurrency Control	/ 15
5 – Crash Recovery	/ 14
6 – Distributed databases	/ 13
7 – Data warehousing	/ 14
8 – Data mining	/ 10
Total	/ 100

1 Evaluating Relational Operators — 14 points

A. (2 points) What is a *left deep plan*?

B. (2 points) What is an *index only scan*?

C. (2 points) Consider the following relational schema of an Employee relation:

$$Employee(\underline{eid : \texttt{int}}, ename : \texttt{string}, address : \texttt{string}, age : \texttt{real})$$

For each of the following indexes, say whether the index matches the given selection conditions. List the primary conjuncts for every match. That is, say **yes** or **no**, and give the primary conjunct.

(1) A B+ tree index on the search key $\langle Employee.eid, Employee.age\rangle$

(a) $\sigma_{Employee.age=20 \wedge Employee.eid<40}(Employee)$

(2) A hash index on the search key $\langle Employee.eid, Employee.age\rangle$

(b) $\sigma_{Employee.eid=20{,}000}(Employee)$

D. (8 points) Answer **ONLY TWO** of the following five questions:

- (4 points) Explain how the **page** nested loop join algorithm works.
- (4 points) Explain how the **sort-merge** join algorithm works.
- (4 points) Explain how the double buffering algorithm works.
- (4 points) Explain how the algorithm for the union operator based on sorting works.
- (4 points) Explain how the hash-based algorithm for the aggregate operators used in combination with GROUP BY works.

2 External Sorting — 10 points

Suppose you have a file with 9 pages along with a set of buffer pages. The 9 pages are laid out on disk as follows:

5,6	8, 4	11, 6	10, 9	7, 8	5, 3	4, 5	7,3	4

Assuming that our most general external sorting algorithm is used, answer the following questions:

1 (6 points) Show a three-way merge sort of this nine page file.

2 (4 points) How many runs are necessary to completely sort the file?

3 Query Optimization — 10 points

Part A —2 points What is a *reduction factor* of a selection condition?

Part B —8 points Consider the following relational schema:

$$Player(\underline{pid : \texttt{int}}, pname : \texttt{string}, age : \texttt{real}, nationality : \texttt{string})$$
$$PlaysIn(\underline{pid : \texttt{int}, tid : \texttt{integer}}, years : \texttt{real})$$
$$Team(\underline{tid : \texttt{int}}, tname : \texttt{string}, owner : \texttt{string})$$

Moreover, consider the following SQL query:

```
SELECT P.pname, T.tname
FROM Player P, Team T, PlaysIn I
WHERE P.pid = I.pid AND T.tid = I.tid
             AND P.nationality = 'Greek'
             AND T.owner ='tiller'
```

Suppose that we have the following indexes: a B+ tree index on the *pid* attribute of the *Player* relation, a B+ tree index on the *pid* column of the *PlaysIn* relation, and a hash index on the *owner* column of the *Team* relation.

Draw a query evaluation plan for the above query and motivate your choice of operator precedence.

4 Concurrency Control — 15 points

Part A —3 points Why are two-phase locking protocols called “two-phase”?

Part B —8 points Consider the following schedule considering of three transactions T1, T2 and T3:

$$S_6 = R_1(X), R_2(Z), R_1(Z), R_3(X), R_3(Y), W_1(X), C_1, W_3(Y), C_3, R_2(Y), W_2(Z), W_2(Y), C_2$$

Answer the questions below:

A. (4 points) Draw the serializability graphs of the schedule.

B. (2 points) Determine whether the schedule is conflict serializable. If so, give an equivalent serial schedule.

C. (2 points) Determine whether the schedule is recoverable. Motivate your answer.

Part C —4 points Consider the following schedule which, in addition to read and write actions, includes locking actions S(X) and X(A).

$$\mathtt{S}_5 = \mathtt{S}_1(\mathtt{U}), \mathtt{R}_1(\mathtt{U}), \mathtt{X}_2(\mathtt{Y}), \mathtt{W}_2(\mathtt{Y}), \mathtt{S}_1(\mathtt{Y}), \mathtt{S}_3(\mathtt{Z}), \mathtt{R}_3(\mathtt{Z}), \mathtt{X}_2(\mathtt{Z}), \mathtt{X}_4(\mathtt{Y}), \mathtt{X}_3(\mathtt{Y})$$

Draw the Wait-for graph for this schedule and indicate if there is any deadlock or not.

5 Crash Recovery — 14 points

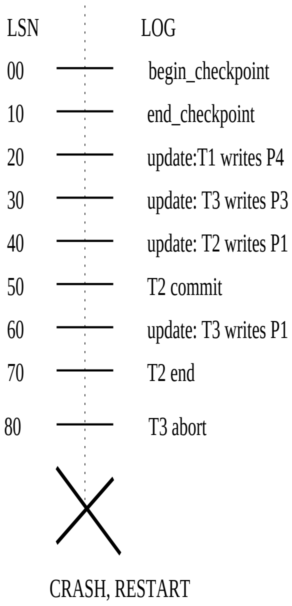
Part A —6 points

A. (2 points) Describe the steal and no-force policy.

B. (4 points) Describe the four properties of transactions that a DBMS much guarantee to maintain data in the face of concurrent access and system failures.

Part B —8 points

Consider the execution shown in the following figure.



LSN	LOG
00	begin_checkpoint
10	end_checkpoint
20	update:T1 writes P4
30	update: T3 writes P3
40	update: T2 writes P1
50	T2 commit
60	update: T3 writes P1
70	T2 end
80	T3 abort

Suppose the system crashes during recovery after writing two log records to stable storage and again after writing another two log records.

Answer the following questions.

A. (3 points) Show what steps are performed during the Analysis phase.

B. (3 points) Show what steps are performed during the Redo phase.

C. (2 points) Show what steps are performed during the Undo phase.

6 Distributed databases — 13 points

Part A —4 points Explain the difference between synchronous and asynchronous replication.

Part B —9 points

Consider a distributed DBMS consisting of the following two tables, involving Customers who rent vacation Homes throughout the Caribbean. The information about all the Customers is stored in Ottawa and all of the RentalHome information is stored in New Providence (in the Bahamas).

- Customer(Cid: integer, Rentid: integer, Income: real)
- RentalHome(Rentid: integer, OwnerId: integer, RentAmount: integer)

The Customer relation contains 100,000 pages and the RentalHome relations contains 5,000 pages. There are no join indexes and a sort-merge join is used locally.

Answer the following questions.

A. (3 points) Consider a query to select all the details of customers who are also owners of homes. (That is, Customer.Cid = RentalHome.OwnerId). This query is posed in Paris, France and you are told that 1 percent of customers are managers. Which query evaluation plan would you use, in order to minimizes shipping costs? Explain your answer.

B. (4 points) Suppose that all the Customer relation tuples are still stored in Ottawa, but the tuples of Customers with income less than 100,000 are replicated in Paris. (That is, the database is now distributed over three sites.) The locks are managed at the *primary site*, i.e. in Ottawa. Explain what locks are set (and at which site) for a query issued at New York City, which wants to read a page of Customer tuples with income less than 50,000.

C. (2 points) Explain what a *phantom deadlock* is and give an example of a situation in the Home Rental distributed database where it may occur.

7 Data warehousing — 14 points

Part A —2 points Explain what view materialization is.

Part B —12 points The School of IT and Engineering' s (SITE) director keeps track of the lab usages, measured by the number of students who uses the labs. This function is very important for budgeting purposes.

You are asked to develop a small data warehouse to keep track of the lab usage statistics. The main requirements are:

- Show the total number of users by different time periods.
- Show usage numbers by time period, by degree and students classification (undergraduate or graduate).
- Compare the usage for different degrees and semesters.

For each student, we keep information including her student identifier, her password (encrypted), name, address, email, together with the classes she is enrolled in, her major, her degree, and her year level. Each time a student logs onto a computer, or disconnects, her student identifier is captured together with the exact time, and this is placed in a log. For each class, we also keep information including the professor, the semester begin and end dates, the topic, as well as the software requirements.

A. (6 points) Draw the lab usage star schema (dimensional model).

B. (3 points) Give an example of an attribute hierarchy that will be used in your data warehouse.

C. (3 points) Give an example of an OLAP query which could be executed against the lab usage data warehouse.

8 Data mining — 10 points

Part A —2 points Explain what **market basket analysis** is and give an example of a method to count co-occurrences in data.

Part B —8 points You are interested in construct a model to determine whether you will be able to go to the beach on a particular day. To this end, you have collected the following data about past weather conditions, together with your past decision (yes or no).

Temperature	Wind	Snow	Sun	Decision
-20	0	no	yes	no
25	5	no	yes	yes
14	2	no	no	yes
13	0	yes	yes	yes
3	30	no	yes	no

Complete **ONLY ONE** of the following two questions.

A. Describe the decision tree induction algorithm you will use to build this model and give an example of a possible decision tree which may be constructed. Note that your decision tree should have at least two levels of internal nodes.

B. Describe an algorithm to find association rules and give an example of (at least three) possible association rules which may be constructed.