CSci 4707
Homework 3
Fall 2015

Chapter 8 and 12
Due Tuesday, 11/17/2015 13:00

A. (5 Points) Submission Guidelines
- All homeworks **must be typed**
- All homeworks due at the **beginning of the class in paper**
- **No partial credit awarded for this section**

B. (25 Points) Chapter 8

1. **(10 Points, 2.5 Points each)** Assume you have a buffer with a maximum capacity of 5 pages. Below is a queue of the **current** state of the buffer with **its current pin count and the timestamp when pin counts becomes 0**.

A:1:NaN	B:0:3	C:0:2	D:0:4	E:0:1

All pages were inserted sequentially, i.e., Page A comes before Page B, etc. and a timestamp of 0 came first before the timestamp of 1 etc.). For example, the pin count of A is 1, thus the timestamp when the pin count of A becomes 0 is NaN. Furthermore, when comparing B and C, the pin count of C becomes 0 first since the value of the timestamp is 2 (compared to 3 in B). Now, there are 4 pages (F, G, H, and I) need to be inserted into the buffer and none of the page will have a pin count changed to 0 during the insertion. For each buffer management policy below, show the result of the buffer after the insertion. Just show the Page name on each buffer slot. Remember the location of each page inside the buffer is important.

   a. **(2.5 Points)** FIFO (First In First Out)

    b. **(2.5 Points)** MRU (Most Recently Used)

    c. **(2.5 Points)** LRU (Least Recently Used)

    d. **(2.5 Points)** Clock (The clock is currently pointing to C and it goes to the right. So start from C.)

2. **(5 Points)** The file systems of modern operating systems also buffer data from disk to in-memory for fast access, and they have very sophisticated replacement strategies. Why does a database system manage its own buffer pool rather than letting the operating system manage the buffer? Give two reasons in **one sentence each**.

3. **(5 Points, 2.5 Points each)** Consider a file that is stored on disk. What buffer manager replacement policies (FIFO, LRU, MRU, or Clock) would you choose for each of the following file access patterns? Clearly state the reason for your choice; answers without a correct reason will receive zero points.

    a. **(2.5 Points)** The entire file is sequentially scanned many times.

    b. **(2.5 Points)** File pages are accessed according to a Zipfian distribution, i.e., popular items are requested very frequently, and unpopular items are requested rarely.

4. **(5 Points, 2.5 Points each)** Answer the following questions:

    a. **(2.5 Points)** When will equality search is better using Unclustered B+Tree rather than Sorted?

    b. **(2.5 Points)** Is it possible to have an Unclustered B+Tree index with Alternative 1? If yes, what is the advantage of using Alternative 1 compared to Alternative 2 or 3? If no, why not?

C. (70 Points) Chapter 12

1. **(20 Points, 4 Points each)** Consider the Employee schema below:
   Employee(<u>eid:integer</u>, ename:string, age:integer, resdience_state:string, category:string)
   For each of the following index structure, can the index be used to answer the query stated for the index? If no, why?
   a. Hash Index on <ename, age>
      i.   **(4 Points)** Employee.age = 30
      ii. **(4 Points)** Employee.ename = "Jacob"
      iii. **(4 Points)** Employee.age = 30 AND Employee.ename = "Jacob"
   b. B+Tree index on <eid, ename>
      i.   **(4 Points)** Employee.ename = "Jacob"
      ii. **(4 Points)** Employee.eid = "115"

2. **(20 Points, 5 Points each)** Consider the Employee schema from Question C1: Suppose that the following indexes, all using Alternative (2) from your textbook for data entries, exist: a hash index on eid, an unclustered B+Tree index on age, a hash index on residence_state and a clustered B+Tree index on <residence_state, age>. The average age of all employees in the company is 35 and 90% of the employees live in California (State="CA"). For each of the following query, how are you going to fetch the query?
   a. **(5 Points)** age >10
   b. **(5 Points)** residence_state ="CA"
   c. **(5 Points)** age > 5 AND category="Manager"
   d. **(5 Points)** residence_state="NY" AND age > 5 AND category="Intern"

3. **(10 Points)** Suppose that, for the above selection condition (C2.b), you want to compute the average age for each residence_state (i.e., group by

state). For selection condition (b), describe the least expensive evaluation method.

4. **(20 Points, 5 Points each)** Suppose relation R has 1000 records and relation S has 5000 records. One page can contain 10 R records or 500 S records. What are the cost of the following join operation between R and S?
   a. **(5 Points)** Index Nested Loop Join with Clustered B+Tree Index on S (R is the outer).
   b. **(5 Points)** Index Nested Loop Join with Clustered Hash Index on R (S is the outer).
   c. **(5 Points)** Index Nested Loop Join with Unclustered B+Tree Index on S (R is the outer).
   d. **(5 Points)** Tuple Nested Loop Join (S is the outer).