CSE 132C Spring 2022 Midterm Exam

Full Name:

Student ID:

INSTRUCTIONS

1. You have 50 minutes to complete this exam.

2. You can have up to one letter/A4-sized sheet of notes, formulae, etc. Apart from this, the exam is closed book/notes/electronics/peers.

3. Please wait until being told to start reading and working on the exam.

4. If you think a question is ambiguous, write down your assumptions, argue that they are reasonable, and then work on the problem using those assumptions.

5. Please ensure that your writing is clear and legible!

<table>
<thead>
<tr>
<th></th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Q 2</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Q 3</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Q 4</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>
I think we should build an SQL database.

Un-oh.

Does he understand what he said or is it something he saw in a trade magazine ad?

What color do you want that database?

I think mauve has the most RAM.
Q 1. [12pts] For each question below, select the right option (only one is correct).

1. Which of the following tells us if a page in the buffer pool is edited by a query?
   (A) PinCount   (B) DirtyBit   (C) Wait queue
   (D) Buffer replacement policy   (E) Number of buffer frames

2. It is not necessary that star joins are always also the following form of joins.
   (A) Snowflake Joins   (B) Equi Joins
   (C) Theta Joins   (D) Inner Joins   (E) Natural Joins

3. Which of the following index alternatives causes the size of a data entry to depend on the arity of the relation being indexed?
   (A) AltRecord   (B) AltRID   (C) AltRIDList   (D) None of ABC   (E) All of ABC

4. Which of the following storage hardware is byte addressable?
   (A) Magnetic hard disk   (B) Flash / SSD   (C) Tape
   (D) Non-volatile RAM   (E) All of these

5. Consider the relation schema $Customers$(CustomerID, Name, Age, Zipcode, SSN) with primary key CustomerID and alternate key SSN. Which of the following indexes is not a unique index?
   (A) Clustered B+ tree on (SSN, Name)   (B) Clustered B+ tree on (Age, CustomerID)
   (C) Unclustered B+ tree on (Name, Age)   (D) Unclustered B+ tree on (Age, SSN)
   (E) Clustered B+ tree on (CustomerID, Age, SSN)

6. Consider the same relation schema as the previous question. Which of the following indexes will certainly match the following selection query: $\sigma_{Age >= 40}(Customers)$?
   (A) Clustered B+ tree on (SSN, Name)   (B) Clustered B+ tree on (Age, CustomerID)
   (C) Unclustered B+ tree on (Name, Age)   (D) Unclustered B+ tree on (Age, SSN)
   (E) Clustered B+ tree on (CustomerID, Age, SSN)
Q 2. [15pts] Suppose we are sorting a relation with 100 million pages and we have 100
buffer pages for the external merge sort (EMS). A "pass" over the relation is defined as one
read and write of the whole file. In all of the following, you have to include both the sort
and merge phases. Clearly circle the correct answer for each of the following questions.

1. [2pts] How many passes will a naive 2-way EMS perform? (Hint: $2^{10} \approx 10^3$)
   (a) 25        (b) 26        (c) 27        (d) 28        (e) 29        (f) 30

2. [2pts] How many passes will a standard multi-way EMS perform, assuming we do
   not use any of the three improvements discussed in class?
   (a) 3       (b) 4       (c) 5       (d) 6       (e) 7       (f) 8

3. [2pts] How many passes will a multi-way EMS perform, assuming we use replace-
   ment sort for internal sorting?
   (a) 3       (b) 4       (c) 5       (d) 6       (e) 7       (f) 8

4. [3pts] How many passes will a multi-way EMS perform, assuming we use replace-
   ment sort for internal sorting along with double buffering but no blocked I/O?
   (a) 3       (b) 4       (c) 5       (d) 6       (e) 7       (f) 8

5. [3pts] How many passes will a multi-way EMS perform, assuming we use replace-
   ment sort for internal sorting along with blocked I/O with block sizes of 10 pages
   but no double buffering?
   (a) 3       (b) 4       (c) 5       (d) 6       (e) 7       (f) 8

6. [4pts] Which among the following numbers of buffer pages is the lowest that will
   ensure EMS finishes in just 2 passes? Use any optimization discussed in class that
can lower the number of passes.
   (a) 100     (b) 1,000     (c) 10,000     (d) 100,000     (e) 1 million     (f) 10 million
Q 3. [12pts] Consider the following extendible hash index with 4 slots per bucket.

![Index Diagram]

1. [5pts] Draw the index after the following sequence of operations are completed:
   delete 10*, insert 42*.
2. **[7pts]** Now draw the index after the following additional sequence of operations are completed: insert 32*, delete 9*. Note that these operations are performed after the previous question’s operations on the original index shown.
Q 4. [11pts] You are given the following simplified relational database for shopping. We use aliases $P$ for Products, $G$ for Category, $C$ for Customers, and $R$ for Purchases.

$P$: Products (PID, PName, Vendor, Price)
$G$: Categories (PID, Category)
$C$: Customers (CID, CName, Zipcode, Age)
$R$: Purchases (CID, PID, Date, Quantity)

Products.PID and Customers.CID are primary keys in their respective tables. In Categories, PID is a foreign key referring to Products.PID. In Purchases, CID and PID are foreign keys referring to Customers.CID and Products.PID, respectively. Age is in years. All relations are assumed to be sets of tuples (no bag semantics).

1. [5pts] Write a relational algebra query to answer the following: Get the set of vendors of products bought by customers of age at least 18 years.

2. [6pts] Write a relational algebra query to answer the following: Get the details of all products in categories that the vendor Nestle does not have products in.