CSIT 5300 - Spring 2017 - HKUST
Assignment 3

Deadline: Wednesday, April 26, 2017, 10:20pm
Submission: Please bring a hard copy of your solution during the lecture

<table>
<thead>
<tr>
<th>Student Name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ID</td>
<td></td>
</tr>
<tr>
<td>Email Address</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td></td>
</tr>
</tbody>
</table>

For T.A. use only

<table>
<thead>
<tr>
<th>Problem</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/ 25</td>
</tr>
<tr>
<td>2</td>
<td>/ 25</td>
</tr>
<tr>
<td>3</td>
<td>/ 25</td>
</tr>
<tr>
<td>4</td>
<td>/ 25</td>
</tr>
<tr>
<td>Total</td>
<td>/ 100</td>
</tr>
</tbody>
</table>
Problem 1 [25 points] Functional Dependencies

(a) [8 points] Consider relation $R_1(A,B,C,D)$ with functional dependencies $F_1 = \{AB \rightarrow C, BC \rightarrow D, CD \rightarrow A, AD \rightarrow B\}$. Identify the strongest Normal form. Explain your answer.

Answer:

(b) [8 points] Consider relation $R_2(A,B,C,D,E)$ with functional dependencies $F_2 = \{AB \rightarrow C, DE \rightarrow C, B \rightarrow D\}$. Give a lossless 3NF decomposition. Explain how you derived your decomposition.

Answer:
(c) [9 points] Consider relation $R_3(A, B, C)$ with functional dependencies $F_3 = \{A \rightarrow BC, B \rightarrow AC, C \rightarrow AB\}$. Find the canonical cover of $F_3$.

**Answer:**
Problem 2 [25 points] B⁺-Trees
Consider a relation with the following schema (the primary key is underlined):

\[
\text{Movie}(id, \text{title}, \text{year}, \text{producerId})
\]

- The size of \( id \) is 4 bytes, \( \text{title} \) is 30 bytes, \( \text{year} \) is 4 bytes, and \( \text{producerId} \) is 4 bytes.
- The size of a pointer is 4 bytes.
- The Movie relation contains 100,000 records.
- Each page has size 1,000 bytes.
- 1,000 movies are produced per year.
- Every producer has produced 500 movies.

(a) [8 points] We construct a B⁺-Tree on \( id \). Suppose that this is a \textbf{primary} (i.e., clustering) index, and that it has 3 levels. What is the cost for answering the following query (in terms of page accesses, without considering the cost of writing the result)?

\[
\text{SELECT title FROM Movie WHERE id = 101}
\]

Answer:
(b) [8 points] We construct a B+\(^{+}\)-Tree on \textit{year}. Suppose that this is a \textbf{primary} (i.e., clustering) index, and that it has 3 levels. What is the cost for answering the following query (in terms of page accesses, without considering the cost of writing the result)?

\begin{verbatim}
SELECT title FROM Movie WHERE year > 2000 AND year < 2006
\end{verbatim}

\textbf{Answer:}
(c) [9 points] We construct a B*-Tree on producerId. Suppose that this is a dense, secondary (i.e., non-clustering) index, and that it has 3 levels. What is the cost for answering the following query (in terms of page accesses, without considering the cost of writing the result)? Assume an 80% page utilization in the index. Also assume that we use the multiple index entries option for storing entries with the same search key (see Option 2 in slide 12 of L07_indexing_intro.pdf).

\[
\text{SELECT id FROM Movie WHERE producerId = 5}
\]

Answer:
Problem 3 [25 points] Extendible Hashing

We consider a hash index using the extendible hashing approach. We assume that each bucket contains 2 entries. In the first four questions below, you are given a figure depicting an instance of the index. For each entry, we illustrate only its search key in both binary and decimal form, omitting the pointers for clarity. The hash function simply maps the records to the buckets, using the rightmost bits of their binary representation (similar to what we have seen in L08_bplus_dynam_hashing.pdf). You are required to draw a figure showing the resulting index contents (i.e., directory, bucket contents, local and global depths) after inserting the given new entries.

(a) [5 points] You are given the following index instance. What is the result after the insertion of a record with key 17 (10001)?

Answer:
(b) [5 points] You are given the following index instance. What is the result after the insertion of two records, one with key 17 (10001) and one with 23 (10111)?

Answer:
(c) [5 points] You are given the following index instance. What is the result after the insertion of a record with key 11 (1011)?

Answer:
(d) [5 points] You are given the following index instance. What is the result after the insertion of two records, both with key 4 (100)? Denote the first new entry as 4*(100), and the second as 4**(100).

Answer:

(e) [5 points] Explain briefly why it is undesirable to have overflow buckets in the hash index.

Answer:
**Problem 4** [25 points] Sorting and Join Algorithms

Consider the following two relations:

- \( R(A, B) \): 15,000 records
- \( S(A, C) \): 40,000 records

Each page is 1000 bytes long. We use a memory buffer of 10 pages. We also assume that all attribute values are 10 bytes long.

(a) [12 points] Demonstrate the external sorting algorithm for sorting \( R \) on \( A \), and explain the total cost (also considering the cost for writing the result back to the disk).

**Answer:**
(b) [13 points] Consider the following SQL query:

\[
\text{SELECT } R.A, B, C \text{ FROM } R, S \text{ WHERE } R.A = S.A
\]

Suppose that we process the SQL query with the merge-join algorithm, assuming that \( R \) and \( S \) are not sorted on any attribute. What is the total processing cost?

Answer: