Student ID:
Student Name:

CISC 3005 Advanced Database Systems
Assignment #1

Instructor: Dingqi Yang

Due: 2025-02-16 @ 23:59

Remind to explain and justify your answers to all questions, including multiple choice questions.

Question 1: Storage Models

Consider a database with a single table C(conference_id, conference_name, host_id, duration, number_of_sessions), where conference_id is the primary key, and all attributes have the same fixed width. Suppose C has 10,000 tuples that fit into 50 pages. You should make the following assumptions:

- The DBMS does not have any additional meta-data or any additional storage overhead for the table (e.g., page headers, tuple headers). C does not have any indexes (including for primary key conference_id).
- None of C's pages are already in memory before a query starts. The DBMS can store an infinite number of pages in memory.
- The tuples of C can be in any order (keep this in mind when computing the minimum versus maximum number of pages that the DBMS will potentially have to read and think of all possible orderings).
- There will always be at least one result from any query.

(a) Consider the following query:

```
SELECT MAX(number_of_sessions) FROM C
WHERE duration > 1440 AND host_id == 5;
```

(i) **Suppose the DBMS uses the decomposition storage model (DSM) with implicit offsets. How many pages will the DBMS potentially have to read from disk to answer this query? Please consider best and worst-case scenarios. (10%)**

(ii) **Suppose the DBMS uses the N-ary storage model (NSM). How many pages will the DBMS potentially have to read from disk to answer this query? (10%)**

(b) Now consider the following query:

```
SELECT conference_name, host_id FROM C
WHERE conference_id = 39 OR conference_id = 49;
```

Suppose the DBMS uses the decomposition storage model (DSM) with implicit offsets.

(i) **What are the minimum and the maximum numbers of pages that the DBMS will potentially have to read from disk to answer this query? (10%)**

Suppose the DBMS uses the N-ary storage model (NSM).

(iii) **What are the minimum and the maximum numbers of pages that the DBMS will potentially have to read from disk to answer this query? (10%)**

(c) Consider the following query:

```
SELECT conference_id FROM C
WHERE duration = (SELECT MIN(duration) FROM C);
```

Suppose the DBMS uses the decomposition storage model (DSM) with implicit offsets.

(i) **What is the number of pages that the DBMS will potentially have to read from disk to answer this query? Please consider the best and worst-case scenarios. (10%)**

Question 2: Slotted Pages and Log-Structured Storage

(a) **Which problems are associated with the slotted-page storage in a database system? Select all that apply. (10%)**

□ Fragmentation

□ Write Amplification

□ Increased Random Writes

□ Increased Random Reads

(b) **Which problems are associated with the log-structured storage in a database system? Select all that apply. (10%)**

- □ Fragmentation
- □ Write Amplification
- □ Increased Random Writes
- □ Increased Random Reads

Question 3: Database Compression

(a) Suppose that the DBMS has a VARCHAR column storing the following values:

```
[Museum of Art, Andy Warhol Museum, Museum of Natural
History, Children's Museum, Solders & Sailors]
```

Which of the following are valid encodings (uint32) for this column under dictionary compression as discussed in lecture that will support both point queries and range queries? Select all the valid encodings. (10%)

□ [3,1,4,2,5]

□ [10,20,30,40,50]

□ [32,15,92,31,196]

□ [31,15,92,32,196]

□ [30,10,40,20,50]

(b) **Suppose the DBMS wants to compress a table R(a) using columnar compression. Which of the following compression schemes will benefit (in terms of compression efficacy) from sorting the table before compressing column a? Select all that apply. (10%)**

□ Run-length Encoding

□ Bit-packing Encoding

□ Bitmap Encoding

□ Delta Encoding

□ Dictionary Encoding

(c) **Which of the following statements are true? (10%)**

- □ Run-length Encoding is effective for compressing any integer column.
- □ Bitmap Encoding on high cardinality columns hurts inserts and updates.
- □ Delta Encoding is good at compressing large text values.
- □ For point lookup-only workload, order-preserving dictionary encoding is unnecessary.