CARNEGIE MELLON UNIVERSITY
COMPUTER SCIENCE DEPARTMENT
15-445/645 – DATABASE SYSTEMS (SPRING 2025)
PROF. JIGNESH PATEL

Homework #2 (by Joe & Aditya)

Due: **Sunday Feb 09, 2025 @ 11:59pm**

IMPORTANT:

- Enter all of your answers into **Gradescope by 11:59pm on Sunday Feb 09, 2025.**
- **Plagiarism**: Homework may be discussed with other students, but all homework is to be completed **individually**.

For your information:

- Graded out of **100** points; **3** questions total
- Rough time estimate: ≈4-6 hours (1-1.5 hours for each question)

Revision : 2025/01/30 12:57

Question	Points	Score
Slotted Pages and Log-Structured	30	
Storage Models	35	
Database Compression	35	
Total:	100	

Question 1: Slotted Pages and Log-Structured . . . . . [30 points]

(a) **[10 points]** Which problems are associated with the *slotted-page storage* in a database system? Select all that apply.
- ☐ Increased Random Writes
- ☐ Write Amplification
- ☐ Fragmentation
- ☐ Increased Random Reads
- ☐ None of the above

(b) **[10 points]** Which problems are associated with the *log-structured storage* in a database system? Select all that apply.
- ☐ Write Amplification
- ☐ Increased Random Writes
- ☐ Increased Random Reads
- ☐ Fragmentation
- ☐ None of the above

(c) **[10 points]** You are asked to compare *log-structured storage* to *slotted-page storage* for a new system. Ignore any indexes and overhead from metadata. Select all true statements.
- ☐ Log-structured storage requires less disk space.
- ☐ Only log-structured storage supports variable length tuples.
- ☐ For an append-only workload, both achieve comparable performance.
- ☐ After lots of insert/update/deletes, only log-structured benefits from maintenance.
- ☐ Log-structured storage is not suitable for systems with limited memory.
- ☐ None of the above are true.

Question 2: Storage Models . . . . . [35 points]

Consider a database with a single table `E(player_id, games_played, room_id, total points)`, where `player_id` is the *primary key*, and all attributes are the same fixed width. Suppose `E` has 20,000 tuples that fit into 100 pages. You should ignore any additional storage overhead for the table (e.g., page headers, tuple headers). Additionally, you should make the following assumptions:

- The DBMS does *not* have any additional meta-data.
- `E` does *not* have any indexes (including for primary key `player_id`).
- None of `E`'s pages are already in memory. The DBMS can store an infinite number of pages in memory.
- Content-wise, the tuples of `E` will always make each query run the longest possible and do the most page accesses.
- The tuples of `E` can be in any order (keep this in mind when computing *minimum* versus *maximum* number of pages that the DBMS will potentially have to read and think of all possible orderings)

(a) Consider the following query:

```
SELECT MAX(total points) FROM E
    WHERE games_played < 445 AND room_id == 15645 ;
```

i. **[5 points]** 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?
Be sure to keep in mind the assumption about the contents of `E`.
☐ 1-25 ☐ 26-50 ☐ 51-75 ☐ 76-100 ☐ ≥ 101 ☐ Not possible to determine

ii. **[5 points]** 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?
Be sure to keep in mind the assumption about the contents of `E`.
☐ 1-40 ☐ 41-60 ☐ 61-80 ☐ 81-100 ☐ ≥ 101 ☐ Not possible to determine

(b) Now consider the following query:

```
SELECT total points, games_played, room_id FROM E
   WHERE player_id = 445 OR player_id = 645 OR player_id = 799
```

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

α) **[5 points]** What is the *minimum* number of pages that the DBMS will potentially have to read from disk to answer this query?
☐ 1-3 ☐ 4-6 ☐ 7-9 ☐ 10-100 ☐ ≥ 101 ☐ Not possible to determine

β) **[5 points]** What is the *maximum* number of pages that the DBMS will potentially have to read from disk to answer this query?
☐ 1-20 ☐ 21-40 ☐ 41-60 ☐ 61-80 ☐ 81-100 ☐ ≥ 101
☐ Not possible to determine

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

α) **[5 points]** What is the *minimum* number of pages that the DBMS will potentially have to read from disk to answer this query?
☐ 1 ☐ 2-3 ☐ 4-6 ☐ 7-9 ☐ 10-100 ☐ ≥ 101 ☐ Not possible to determine

β) **[5 points]** What is the *maximum* number of pages that the DBMS will potentially have to read from disk to answer this query?
☐ 1 ☐ 2-3 ☐ 4-6 ☐ 7-9 ☐ 10-100 ☐ ≥ 101 ☐ Not possible to determine

(c) Finally consider the following query:

```
SELECT player_id FROM E
   WHERE total points = (SELECT MAX(total points) FROM E);
```

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

i. **[5 points]** What is the *minimum* number of pages that the DBMS will potentially have to **read from disk** to answer this query?
☐ 1-20 ☐ 21-40 ☐ 41-60 ☐ 61-80 ☐ 81-100 ☐ ≥ 101 ☐ Not possible to determine

Question 3: Database Compression . . . . . [35 points]

(a) **[5 points]** Suppose that the DBMS has a `VARCHAR` column storing the following values:

```
[Woody, Buzz Lightyear, Mike Wazowski, Lightning
McQueen, Lightning Storm]
```

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.

☐ `[5,1,4,2,3]`
☐ `[79,12,32,15,33]`
☐ `[79,12,33,15,32]`
☐ `[50,10,40,20,30]`
☐ `[10,20,30,40,50]`

(b) **[15 points]** Suppose the DBMS wants to compresses a table `R(a)` using columnar compression. Which of the following compression schemes **will not benefit** from sorting the table before compressing column `a`? Select **all** that apply.
Hint: "Benefit" means that the efficacy of the compression scheme improves on sorted data. You should not make any assumptions about the column type or its distribution of values.

☐ Run-length Encoding
☐ Bit-packing Encoding
☐ Mostly Encoding
☐ Bitmap Encoding
☐ Delta Encoding
☐ Dictionary Encoding
☐ All of the above will not benefit.

(c) **[15 points]** A colleague approaches with a list of true and false statements about run-length encoding, delta encoding, bitmap encoding, and dictionary encoding. The colleague wants your assistance in identifying the true statements. Select **all** that apply.

☐ 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.
☐ For a heavy update workload, dictionary performs better than delta encoding.
☐ None of the above.