CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5481 Data Engineering
Session	:	Semester A 2020/21
Total time allowed	:	70 minutes (35 minutes / section)

1. This paper consists of **2** sections: Section A and Section B.
2. Answer **ALL** questions in both sections.
3. Specify the Section and Question number clearly for **EACH** answer in the answer script.
4. Submit **ONE** pdf file for **EACH** section to Canvas.
5. Use your Student ID and Section number to name the pdf file, e.g., "51234567A.pdf" and "51234567B.pdf".

*This is an **open-book** quiz.*
***NO** access to the Internet, except for the operation of the quiz.*
Candidates are allowed to use an approved calculator.

Copy-and-paste the following academic honesty pledge on the first page of the Section A answer script.

"I pledge that the answers in this examination/quiz are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

- ***I will not plagiarize (copy without citation) from any source;***
- ***I will not communicate or attempt to communicate with any other person during the examination/quiz; neither will I give or attempt to give assistance to another student taking the examination/quiz; and***
- ***I will use only approved devices (e.g., calculators) and/or approved device models.***
- ***I understand that any act of academic dishonesty can lead to disciplinary action."***

Write the following together with your student ID and name to reaffirm the academic honesty pledge onto the first page of the Section A answer script.

"I pledge to follow the Rules on Academic Honesty and understand that violations may lead to severe penalties."

Student ID: _____

Student Name: _____

Section A

Query processing and optimization [25 marks]

Consider the following relations, where the keys are underlined:

ENGINEER (ID, Name)
PROJECT (PID, ICEngID)

The ICEngID attribute in PROJECT is the ID of the engineer who is in charge of the project and PID is the ID of the project.

Consider the following query.

```
SELECT    *
FROM      ENGINEER E, PROJECT P
WHERE     E.ID=P.ICEngID
```

Given the following statistics and indices:

- number of tuples in ENGINEER: 1,600
- number of tuples in PROJECT: 3,200
- size of a tuple in ENGINEER: 50 bytes
- size of a tuple in PROJECT: 80 bytes
- disk block size: 512 bytes
- tuples do not span across blocks
- V(ICEngID, PROJECT) = V(ID, ENGINEER) = 1,600
- 3-level B^+-tree primary index on ID for ENGINEER
- 4-level B^+-tree primary index on PID for PROJECT
- 3-level B^+-tree secondary index on ICEngID for PROJECT

a) Estimate the number of output tuples for the query. Explain.

[3 marks]

b) Draw a fully annotated ***evaluation plan*** if the query is computed with the ***merge-join*** algorithm for the ***worst-case estimate***.

[4 marks]

c) What is the **minimum** amount of memory in number of blocks for the ***worst-case estimate*** of the evaluation plan in part b)?

[1 mark]

d) What is the ***worst-case cost*** in ***number of disk block transfers*** of the evaluation plan in part b)? Show the steps clearly (no steps, no marks).

[11 marks]

e) Is it possible to reduce the ***worst-case cost*** in ***number of disk block transfers*** if the query is computed with the ***indexed nested-loop join*** algorithm instead? Draw a revised evaluation plan and show the steps clearly to support your answer (no steps, no marks).

[6 marks]