CS 2550 – Principles of Database Systems (Spring 2024)
Dept. of Computer Science, University of Pittsburgh

Assignment #4: Query Processing & Optimization

Release: April 10, 2024 Due: 8:00PM, April 22, 2024

Goal
The goal of this assignment is to better understand the query evaluation process.

Description
The following is a fragment of the university database schema.

```
Employee (EmployeeID, First, Last, Gender, DOB, Address, Phone, SNO, Salary)
PK(EmployeeID)
FK(SNO) --> Studio(StudioNum)

Studio (StudioNum, StudioName, Manager, Budget)
PK(StudioNum)
FK(Manager) --> Employee(EmployeeID)

Movie (movieID, producedBy, title, Budget, filmedAt)
PK(movieID)
FK(producedBy) --> Studio(StudioNum)
```

Consider four possible organizations of the tables `Employee(E)-Studio(S)-Movie(M)`:

1. Tables E, S and M are heap files & no access methods exist on any of them
2. Tables E and S are sorted files and M is a heap file & no access methods exist on any of them
3. Tables E, S and M are heap files & access methods (hashing) exist only on M
4. Table E is a sorted file and tables S and M are heap files & access methods exist on E and M, one of which is hash and the other is index

Further assume the following statistics:

- Available Cache Memory (CM) to be 22 pages.
- Attributes belonging to the same table all have the same average size.
- When using hash join, the smaller record will be the partitioning part.
- The distribution of values for the attribute filmedAt in the table Movie is: 50% 'PGH', 30% 'PHL' and 20% 'NYC'.
- Movie: $r = 500, bfr = 5, B_{Prj} = 100$.
- Studio: $r = 250, bfr = 5, B_{Dep} = 50$.
- Employee: $r = 1250, bfr = 5, B_{Emp} = 250$.
- Note that r is the total number of tuples in a table, bfr is the blocking-factor, and B is the total number of blocks of a table.

Questions [100 points]

Q1 [20 points] Produce the heuristic query evaluation tree of the following SQL statement assuming no access structures available, and describe in details how the evaluation of the query would be:

```
SELECT S.StudioName, COUNT(M.movieID), SUM(M.Budget)
FROM   EMPLOYEE AS E, Studio AS S, Movie AS M
WHERE  M.producedBy = S.StudioNum AND E.SNO = S.StudioNum
GROUP BY S.StudioName
HAVING COUNT(E.EmployeeID) < 15;
```

State any additional assumptions. (Hint: some operators can be combined into a single algorithm).

Q2 [80 points] Consider the query evaluation tree as discussed in class of the following SQL statement.

```
SELECT M.movieID, M.producedBy, E.First, E.Last, E.Gender
FROM   Employee AS E, Studio AS S, Movie AS M
WHERE  M.producedBy = S.StudioNum AND S.Manager = E.EmployeeID
      AND M.filmedAt='PGH';
```

Based on your pittID, you have two database organizations assigned to you from the above organizations in order to answer this question. For those two organizations, identify sub trees that represent groups of operations which can be executed by a single algorithm. For each such sub tree, there are potentially multiple methods that can be combined to implement it as a single algorithm. Please provide the following:

- The description of one such implementation (steps & data structures) out of all possible ones for each of the above cases.
- The I/O cost of your implementation at each level.
- Compute the total I/O cost.

You are asked to answer the question for two database organizations, those are based on your pittID, as follows:

ddh32, tiw81, yiv161, dhb51, yul251, sml153, nia135, dtk28:
You are required to do the first and second organizations only.

mas937, yux85, anp407, shk148, nat134, aba166, mod53, jeb386:
You are required to do the first and third organizations only.

avm50, bsp22, yuw328, shy158, btl26, shp184, jah292, tbt8:
You are required to do the first and fourth organizations only.

What & How to submit

- You are required to submit exactly one **PDF** file under your pitt_user_name (e.g., pitt01.pdf). In addition to providing the answer, you are expected to: **include your name and pitt ID at the top of the PDF file.**
- After preparing your solution, submit your file (i.e., the prepared PDF) that contains your solution by going to the class' Gradescope by either navigating to the Gradescope Homepage and selecting the course CS 2550 from the Course Dashboard or by clicking the Gradescope Navigation option under our course Canvas page.
- You must submit your assignment before the due date (**8:00 PM**, Apr. 22, 2024). **There are no late submissions.**
- Note that you are required to use a graph tool (such as MS-Word, MS Powerpoint, MS Visio, idraw, draw.io, etc.) to generate your diagrams and the text that explain them. **Handwritten/Hand-drawn diagrams and/or text will not be accepted/graded and will receive a zero. Hand-written and digitized/scanned images will receive a zero (including scanned, photos, or electronically drawn using a smart pen submissions)**.

Academic Honesty

The work in this assignment is to be done *independently*. Discussions with other students on the assignment should be limited to understanding the statement of the problem. Cheating in any way, including giving your work to someone else will result in an **F** for the course and a report to the appropriate University authority.