December 1, 2016
Arnab Chakrabarti, Rihan Hai, Sandra Geisler

Implementation of Databases (WS 16/17)

Exercise 4

Due until December 13, 2016, 2pm.
Please submit your solution *in a single PDF file* before the deadline to the L²P system!
Please submit solutions in groups of three students.

Exercise 4.1 (I/O Costs of Access Paths) (10 pts)

Referring to Slide 31 of Chapter 2, please reason below formulas **in detail**:

1. Why is the cost for an equality selection using a sorted relation $D * \log_2 B$?
2. Why is the cost for a Range Selection using a clustered tree index $D * (\log_G 0.15B + \sharp matchingpages)$?
3. Why is the cost for an equality selection using an unclustered hash index $2D$?
4. Why is the cost for a delete operation using an unclustered tree index $D * (3 + \log_G 0.15B)$?

Base your explanation on the below assumptions from empirical studies:

- In a sorted file, pages are stored sequentially, retrieving a desired page directly only needs one disk I/O.
- In a clustered file, pages are usually 67% full, and the number of physical data pages is 1.5B.
- We omit the time for processing a record in memory (since it is usually negligible compared with the time for reading or writing disk pages)

Exercise 4.2 (Query Optimization) (12 pts)

Given a relational table EMPL(eno,name,salary,marstat,dno) which is stored in an unsorted heap file with 1,000 pages (primary key is eno). Your system should be optimized for the following queries:

1. `Q1: SELECT * FROM EMPL WHERE eno = 4711`
2. `Q2: SELECT name,salary FROM EMPL WHERE salary > 40000 AND salary < 50000`
3. `Q3: SELECT dno, AVG(salary) FROM EMPL WHERE marstat = 'single' GROUP BY dno`

How do you physically organize your database? Which indexes(clustered/unclustered) should be created to optimize the overall performance for all three queries? What are the estimated costs for your solution based on the information on Slide 31 of Chapter 2 (for the fan-out of tree index G we take 100)?

Exercise 4.3 (Datalog) **(8 pts)**

1. Given is the following extensional database:

   - Child(X,Y) : X is child of Y
   - Female(X) : X is a female person

   Define the following relations of the intensional database by specifying appropriate Datalog rules (you may define additional rules for your convenience):

   (a) Cousin (X,Y): X is a cousin of Y

   (b) Nephew (X,Y): X is a nephew of Y

   (c) Uncle (X,Y): X is an uncle of Y

   (d) GreatUncle (X,Y) : X is a great uncle of Y

2. Consider the following Datalog program.

   $F = \{r(3,4), r(5,2), a(5), a(2)\}$

   R :
   $q(X) :- \quad p(X), r(Y,X), b(Y)$
   $b(X) :- \quad r(X,Y), a(X)$
   $p(X) :- \quad a(X), NOTb(X)$

   (a) Define the Herbrand base for the rules R and facts F.

   (b) Is the program stratified? Draw the stratification graph.

   (c) Compute the least fixpoint for the stratified program (stratum by stratum) and the facts F.