# CMPUT 291 - File and Database Management (Fall 2022)



## Exercises on indexing, query processing and DB tuning

**A few exercises on indexing, query processing and DB tuning**

**Q1**. Consider an empty B+-tree where leaves can hold up to 2 entries and non-leaves can hold up to 3 entries.

a) Insert into your index 12 records with keys chosen as random numbers in 1 to 200.
b) Select 4 records randomly from the index and delete them.

**Q2**. Consider an empty extendable hash file where each bucket can hold up to 3 entries.

a) Insert 12 records with keys chosen as random numbers in 1 to 200.
b) select 4 records randomly from the index and delete them.

**Q3**. Consider a disk with 930,408 cylinders, 6 tracks per cylinder, 128 sectors per track and 512 bytes per sector. What is the disk capacity in bytes?

**Q4**. Consider a disk with 128 sectors per track, 512 bytes per sector and average seek time of 8 msec. The disk platters rotate at 5400 rpm. How long does it take to read a block of 20 sectors?

**Q5**. Consider a B+-tree index where each node (leaf and non-leaf) can store up to 200 entries, and you want to insert 10 million index entries to the index.

a) Find the minimum and maximum depth of the tree?
b) Suppose the index is built from scratch using the insert algorithm discussed in class; without worrying about the effects of buffering, estimate the number of I/Os in the worst case.
c) Do the estimation in part (b) but assume the first two levels of the index are stored in main memory.
d) Suppose you can freely sort the index entries! What would be your best algorithm to load the index? Again estimate the number of I/Os without worrying about the effects of buffering.

**Q6**. Consider the following query

*select * from weather where city = "Edmonton" and temp < -30*

Suppose B+-tree indexes are constructed on columns city and temp. Data entries are stored in both indexes as (key,rid,pid) where rid is the record id of a data record with key value k and pid is the address of a page where the record with id rid is stored. Further assume 1% of the tuples satisfy the predicate "city='Edmonton'", 1% of the tuples satisfy the predicate "temp<-30" and 0.01% of the tuples satisfy both predicates.

a) What would be an efficient algorithm to evaluate the query if we know that the index on city is clustered?
b) What would be an efficient algorithm to evaluate the query if we know that none of the indexes are clustered?
c) Suppose each page of the index (leaf and non-leaf) stores up to 200 keys, and each data page stores up to 100 records. Assume both index and data pages are full. If N denotes the number of data records, estimate the number of page accesses for both algorithms given in (a) and (b).

**Q7**. Consider the query in Q6, and suppose the table weather has N rows. Many database systems keep the number of distinct values for each column and use it in their cost estimation.
(a) Suppose the number of distinct values in column city is 40. Assuming uniformity, how many rows are expected to satisfy the condition city = "Edmonton"?

(b) Suppose the minimum and the maximum values of the temp column in table weather are -35 and 30 respectively. Assuming uniformity, how many rows are expected to satisfy the condition temp < -30?

(c) Estimate the number of rows that satisfy both conditions (i.e. city = "Edmonton" and temp < -30). State any assumptions that you make.

(d) Estimate the number of rows that satisfy at least one conditions (i.e. city = "Edmonton" or temp < -30). State any assumptions that you make.

**Q8**. Consider the following tables where sid in transcripts is a foreign key referencing students.

students(sid, name, phone)
transcripts(sid, cid, sem, grade)

Consider the following queries and suppose the students table has M rows and the transcripts table has N rows. Suppose the number of distinct values in column sid of transcripts is 1000, and the number of distinct values in the sem column is 10.

q1.
```
select *
from students s, transcripts t
where s.sid=t.sid;
```

q2.
```
select *
from students s, transcripts t
where s.sid=t.sid and t.sem='F22';
```

(a) Estimate the number of rows returned by q1.

(b) Estimate the number of rows returned by q2.

(c) Assuming each page can hold 100 rows, a main memory buffer that can hold 3 pages and with no index, what is the cost (in terms of the number of I/Os) of evaluating q1?

(d) What is the cost in c if the main memory buffer can hold m pages instead?

(e) What is the cost in c if there is an index on column sid of students?

Last modified: Saturday, 3 December 2022, 10:04 PM