CS 4320/5320: Assignment 2 (70 Points Total)

Note the requirements on the submission format described at the end of this document!

Consider the database schema created by the following commands:

```
CREATE TABLE Sailors (
sid integer PRIMARY KEY, sname varchar(20), rating integer, age real);
CREATE TABLE Boats (
  bid integer PRIMARY KEY, bname varchar(20), color varchar(20));
CREATE TABLE Reserves (
  sid integer, bid integer, day date, primary key (sid, bid, day));
```

In the following, we ask you to "translate" relational algebra queries to SQL queries. We assume that relational algebra queries are executed on three relations: B, R, and S. Relation B has the same columns (i.e., same name, type, constraints, and semantics) as the Boats table created above and contains the same tuples. The same relationship holds between relation R and table Reserves and between relation S and table sailors.

An SQL query is equivalent to a relational algebra query if it always yields exactly the same result (i.e., it yields the same set of tuples), independently of the database content. We assume set semantics for relational algebra (in the sense that duplicates are removed after each operation).

Q1) (10 Points) Find an equivalent SQL query for the following relational algebra query:

$$\Pi_{sname}(S \bowtie (\Pi_{sid}(\sigma_{color='red'}(B \bowtie R)) \cap \Pi_{sid}(\sigma_{color='blue'}(B \bowtie R))))$$

Q2) (10 Points) Find an equivalent SQL query for the following relational algebra query:

$$\Pi_{bid}(B) - \Pi_{bid}(\Pi_{sid,bid}(B \times S) - \Pi_{sid,bid}(R))$$

In the following three questions, we ask you to calculate the cost of retrieving sailors satisfying certain conditions via different access methods. We use the simple cost model presented in class (counting the number of pages read from disk).

We assume that our database contains 50,000 sailors. Storing one character takes one byte, storing one integer takes four bytes (record IDs and pointers to index nodes are integers, too), storing one real number consumes eight bytes. Assume that each disc page stores 8,000 bytes of data (we neglect layout-related meta-data). Further, we assume that ratings are integers between 1 and 10 (inclusive) and the number of sailors having each rating is the same.

Q3) (10 Points) Assume that sailors are sorted by rating - what is the cost of retrieving all sailors with a rating of 10 (using binary search to find the first such sailor)?

Q4) (10 Points) Assume that sailors are indexed by rating via a clustered B+ tree index that contains the data itself (Alternative 1), all tree nodes are stored on disk (also the root). What is the cost of retrieving all sailors with a rating of 10? Note: we require you to calculate the minimal height of the B+ tree index given page size and maximal number of entries per page!

Q5) (10 Points) Assume that sailors are indexed by rating via an unclustered B+ tree index that contains single RIDs (Alternative 2), all tree nodes are stored on disk (also the root). What is the cost of retrieving all sailors with a rating of 10? Note: you need to retrieve all data about each sailor and you need to calculate the precise B+ tree index height as before!

In the following questions, we ask you about the cost of sorting data (number of page I/Os). We assume that the sorting algorithm "General External Merge Sort", as presented in the lecture, is used. We want to sort the Reserves relation, assume that we have 500,000 reservations. Integers take four bytes and dates take eight bytes of storage. Each disk page stores 8,000 bytes of data (neglecting layout-related meta-data).

Q6) (10 Points) What is the cost of sorting with 10 buffer pool pages?

Q7) (10 Points) Assuming that each page I/O takes 1 millisecond, how much buffer pool pages do we need at least to sort all data within at most four seconds?

We use (semi-)automated grading for your submissions so please ensure that your submission follows precisely the format outlined below:

The required format is the following:

- you must submit one .zip file (without any sub-directories) that contains seven text files,
- each text file contains the answer to one question,
- the text files must be named "Q1.txt", …, "Q7.txt",
- for questions one and two, the corresponding text file contains one single SQL query (and nothing else) that must run on the latest Postgres version,
- for the remaining questions, the corresponding text file contains only the plain result number (either a cost value or number of buffer pages for Q7) without units in the first line and a one-paragraph justification of your result afterwards.