

You are free to use any resources you like (e.g., papers, publicly available code, homework solutions found on the Web) in completing this homework, subject to two conditions: (1) All such resources used must be clearly and prominently acknowledged. (2) Everything you write should be original (in your own words) and you should be able to explain everything from scratch.

1. (3×10 points) Consider the following CALC query over $R(A, B, C)$ and $S(B, C, D, E)$:

$$\{ \langle x, y, z \rangle : P, Q, R \mid \\ \exists t_1, t_2 (R(x, t_1, t_2) \wedge \exists t_3 R(t_1, t_3, x)) \\ \wedge \exists t_4 (R(y, x, t_4) \vee R(y, t_4, x)) \\ \wedge S(x, t_5, t_6, z) \\ \wedge \forall t_7 (R(x, t_7, x) \rightarrow \exists t_8, t_9 S(x, x, t_8, t_9)) \}$$

Compute SPJR queries that are equivalent to the above using three methods: (a) the method suggested by the proof of Theorem 5.3.10 of [1]; (b) the method outlined as Algorithm 5.4.8 in [1]; and (c) a method of your choice that results in a query that is likely to be more efficient than those produced by the above methods. Show the intermediate steps for the first two methods. Indicate why the query produced by the third method is likely to be more efficient than the others.

2. (5×8 points) The following questions are about XQuery. When asked for a query, explain briefly why your answer is correct.
 - (a) Write a query to extract from an XML document `library.xml` all `book` and `article` elements that have an `author` child with value “Smith” or “Jones.” The elements in the output should be sorted by publication date, in reverse chronological order. Assume that each `book` and `article` element has a single `pubdate` child with the obvious semantics.
 - (b) Explain what your query for Question 2a returns if each `book` and `article` element in `library.txt` is permitted to have any number (zero or more) of `pubdate` children.
 - (c) Rewrite your query for Question 2a, if needed, so that it returns the books and articles authored by Smith or Jones (as before) but copes with the situation of Question 2b by using the first `pubdate`, in document order, of each item for sorting. Items without `pubdate` children should be listed last.
 - (d) Write a query that outputs a sequence of XML elements of the form

`<author>A<pcount>N</pcount></author>`

where N is the number of books and articles authored by an author named A . The output should be sorted by author names.

(e) Write a query that returns a version of `library.xml` in which each `book` and `article` element is transformed as follows: All (zero or more) `pubdate` subelements (descendants, not necessarily children) of the `book` or `article` are moved to (made children of) a single, newly created `alldates` child of the `book` or `article`. Other than this change, the query result should be identical to `library.xml`. Try not to make additional assumptions on the structure of `library.xml` but if your answer depends on some additional assumptions, be sure to list them clearly.

3. (30 points) In class, we discussed a simple hash-based join algorithm that used a fraction $f \in [0, 1]$ of available main memory for the hash table and the rest as a buffer for recently accessed disk pages. We determined that the algorithm (with $f > 0$) does well compared to the simpler alternative of $f = 0$, assuming uniformly distributed hash values. Repeat the analysis for a 90-10 distribution in which 90% of the items hash to 10% of the hash buckets and the remaining 10% items hash to the other 90% of the hash buckets. Within each category, assume that items are uniformly distributed. You may make additional simplifying assumptions if needed, as long as you list them clearly.

Submission Submit your entire midterm electronically as a single gzipped tar archive. Your answers should be in a single PDF or plain text file. Please check that the PDF file is portable. At the very least, the `gv` program on the CSIC cluster must display it properly. Include a `README` file that describes the contents and their relation to the above questions. Name your file using the scheme `LastnameIJ-mt-NNNN.tgz`, where `NNNN` is a 4-digit integer of your choice, and upload by anonymous FTP to `ftp.cs.umd.edu`, directory `incoming/chaw`.

References

- [1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [2] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice-Hall, 2002.