EPL646 – Advanced Topics in Databases

External Sorting and Evaluation of Relational Operators

http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html

Exercise 13.1

(Exercise 13.1) Suppose you have a file with 10,000 pages and you have three buffer pages. Answer the following questions for each of these scenarios, assuming that our most general external sorting algorithm is used:

I. A file with 10,000 pages and three available buffer pages.
II. A file with 20,000 pages and five available buffer pages.
III. A file with 2,000,000 pages and 17 available buffer pages

Exercise 13.1

(Exercise 13.1) How many runs will you produce in the first pass?:

I. A file with 10,000 pages and three available buffer pages.
II. A file with 20,000 pages and five available buffer pages.
III. A file with 2,000,000 pages and 17 available buffer pages

Exercise 13.1

(Exercise 13.1) How many passes will it take to sort the file completely?:

I. A file with 10,000 pages and three available buffer pages.
II. A file with 20,000 pages and five available buffer pages.
III. A file with 2,000,000 pages and 17 available buffer pages

Exercise 13.1

(Exercise 13.1) What is the total I/O cost of sorting the file?:

I. A file with 10,000 pages and three available buffer pages.
II. A file with 20,000 pages and five available buffer pages.
III. A file with 2,000,000 pages and 17 available buffer pages

Exercise 13.1

(Exercise 13.1) How many buffer pages do you need to sort the file completely in just two passes?:

I. A file with 10,000 pages
II. A file with 20,000 pages
III. A file with 2,000,000 pages

Exercise 14.3

(Exercise 14.3) Consider processing the following SQL projection query:

SELECT DISTINCT E.title, E.ename FROM Executives E

You are given the following information:
Executives has attributes **ename**, **title**, **dname**, and **address**; all are string fields of the same length.
The **ename** attribute is a candidate key.
The relation contains **10,000** pages.
There are **10** buffer pages.

Consider the optimized version of the sorting-based projection algorithm: The initial sorting pass reads the input relation and creates sorted runs of tuples containing only attributes **ename** and **title**. Subsequent merging passes eliminate duplicates while merging the initial runs to obtain a single sorted result (as opposed to doing a separate pass to eliminate duplicates from a sorted result containing duplicates).

Exercise 14.3

(Exercise 14.3) (i) How many sorted runs are produced in the first pass? What is the average length of these runs? (Assume that memory is utilized well and any available optimization to increase run size is used.) What is the I/O cost of this sorting pass?

Exercise 14.3

(Exercise 14.3) (ii) How many additional merge passes are required to compute the final result of the projection query? What is the I/O cost of these additional passes?

Exercise 14.3

(Exercise 14.3) (iii)

a) Suppose that a clustered B+ tree index on title is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?

b) Suppose that a clustered B+ tree index on ename is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?

c) Suppose that a clustered B+ tree index on <ename, title> is available. Is this index likely to offer a cheaper alternative to sorting? Would your answer change if the index were unclustered? Would your answer change if the index were a hash index?

Exercise 14.3

(Exercise 14.3) (iv) Suppose that the query is as follows:

SELECT E.title, E.ename FROM Executives E

That is, you are not required to do duplicate elimination. How would your answers to the previous questions change?

Questions?

http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html