'5 qual questions
nifer Widom

CS

Question 1. Suppose you are given a file of student records, where each record has three fields: ID, course, and grade. You expect to answer many questions of the form: give me all records for ID = x, where x is some constant. Describe the different types of extra mechanisms or structures that could be used so that these types of questions can be answered very efficiently. What are the trade-offs in the different mechanisms/approaches?

Answer 1. (1) Maintain a permanent hash table that maps a record ID to a list of pointers identifying where the records for that ID are located in the file. (2) Use a permanent B-tree indexing structure that, for a given ID, finds the records in the file. (3) Keep the records sorted, either as an optimization to or instead of (1) or (2). If (3) is used instead of (1) or (2), some kind of binary search would be needed. In contrasting (1) and (2), (1) can provide constant time access while (2) requires logarithmic time. However, the performance of (1) can degrade when many new records are added, while the performance of (2) should not.

Question 2: Suppose in a distributed system there is a table R(A,B) at site 1 and a table S(B,C) at site 2. A user at site 1 wishes to get the "join" of tables R and S, i.e., the user wants one record (A,B,C) for every record (A,B) in R and every record (B,C) in S such that the B values match. Assume that the most expensive operation is sending data across the network between sites 1 and 2. Describe two different algorithms for computing the join, and explain in which scenarios th algorithm is preferable.

Answer 2: Algorithm (1) = all of table S is sent from site 2 to site 1; the join operation is performed at site 1. Algorithm (2) = the B values in R are sent from site 1 to site 2; the matching (B,C)s from S are sent from site 2 to site 1; the join is performed at site 1 using the S values received. Algorithm (1) is preferable in the case where most (B,C) values in S are matched in R, since it avoids the extra communication step in which R's B values are transmitted. However, if many S values are not matched in R, and the number of different B values in R is not vastly larger than S, then (2) is preferable since the extra S values are never shipped.