Week 9 Exercises: Query Processing with Relational Operations (part 2)

Exercise 14.1 Briefly answer the following questions:

1. How does hybrid hash join improve on the basic hash join algorithm?
2. Discuss the pros and cons of hash join, sort-merge join, and block nested loops join.
3. If the join condition is not equality, can you use sort-merge join? Can you use hash join? Can you use index nested loops join? Can you use block nested loops join?

Exercise 14.4 Consider the join $R \bowtie_{R.a=S.b} S$, given the following information about the relations to be joined. The cost metric is the number of page I/Os unless otherwise noted, and the cost of writing out the result should be uniformly ignored.

Relation R contains 10,000 tuples and has 10 tuples per page.
Relation S contains 2000 tuples and also has 10 tuples per page.
Attribute b of relation S is the primary key for S.
Both relations are stored as simple heap files.
Neither relation has any indexes built on it.
52 buffer pages are available.

1. What is the cost of joining R and S using a page-oriented simple nested loops join? What is the minimum number of buffer pages required for this cost to remain unchanged?
2. What is the cost of joining R and S using a block nested loops join? What is the minimum number of buffer pages required for this cost to remain unchanged?
3. What is the cost of joining R and S using a hash join? What is the minimum number of buffer pages required for this cost to remain unchanged?
4. What would be the lowest possible I/O cost for joining R and S using *any* join algorithm, and how much buffer space would be needed to achieve this cost? Explain briefly.
5. How many tuples does the join of R and S produce, at most, and how many pages are required to store the result of the join back on disk?

Exercise 14.5 Consider two duplicate-free relations R and S, which share the same schema. R is of size 1000 disk pages; S is of size 100 pages. We wish to evaluate an equijoin of R and S, with predicate R.a = S.a. There are 12 buffer pages available for this operation. Estimate the cost of the three alternative join strategies: block nested loops join, hash join, sort-merge join. The cost metric is the number of page I/Os, and the cost of writing out the result should be uniformly ignored.