Warm-up Exercise

(See exercise sheet. You can start before class.)

(a) ORDER BY(day), external merge sort
   | SELECT(bid=100), on the fly
   | SEQ SCAN(Reserves)

(b) 1000 + 0 + 10 + 10 + 0 = 1020

Goals for Today

• Continue to reason about estimating the result cardinality for selections and joins
  – System R heuristics
  – More advanced: histograms – Lab 3!

• Begin to explore the search space explosion for alternate query plans

• Discuss the influence of index selection

Adminstrivia

• Prof Beth’s Monday office hour updated to Friday 2:45pm – 3:45pm starting after Spring Break
  – 5 “yes”
  – 3 “no”
Result Size Estimation for Joins

- For equi-join of R and S... *range of result sizes (in # of tuples)*?
  - If R and S have no join attribute values in common?
  - If join attributes are a key for S?
    • And if the join attributes are also a foreign key in R?

- General case: join attributes a in common, a key for neither
  - Assumption: the set of distinct R.a values is contained in S.a
  - Idea: each tuple of R has a 1/NKeys(S) chance of joining with a
tuple in S

  \[
  \text{NTuples(R)} \times \text{NTuples(S)} / \text{NKeys(S)}
  \]

  - Reversing above assumption yields

  \[
  \text{Ntuples(S)} \times \text{Ntuples(R)} / \text{Nkeys(R)}
  \]

  (use smaller of two if different)

Histories: Finer-Grained Statistics

- For better RF estimation, many systems use histograms
- Histogram is *approximation of a data distribution*

- Example: ratings of Sailors (40,000 total tuples)

Exercise 6-7 (from last time)

6. Estimate the result cardinality for this SQL query:

   ```
   SELECT *
   FROM Sailors NATURAL JOIN Reserves NATURAL JOIN
   Boats;
   ```

   Answer: number of tuples in Reserves (1000 pages, with 100 tuples/page)

7. Estimate the cost in I/Os of this query plan:

   ![Query Plan Diagram]

   - Join 1 I/O cost = 1000 * 10 * 500
   - Join 2 I/O cost = 1 * 200
   - Total = 6000 + 200 = 6200 I/Os

Equi-width vs. Equi-depth

- Equi-width
  - # values represented by each bucket is the same

- Equi-depth
  - # of records in each bucket is ~same

Which has better selectivity estimate for a very frequent value?
Exercise 2: Histograms

![Diagram](from Lab 3)

Exercise 2: answers

a) 41 values into 10 buckets. 4 in each, last one 5

b) Amount within bucket = 0.25\(*h_b\) tuples → Overall amount = \((0.25\(*h_b\) tuples) / ntups\)

c) 0.25\(*h_b\) tuples + all tuples from buckets \(i > b\) → Divide sum by ntups

d) 0

Creating an Equi-width histogram

- Suppose you want to be able to estimate the selectivity (reduction factor) for this query:

  \[
  \text{SELECT} \ast \text{FROM Sailors S WHERE S.age} = 40 \text{ AND S.rating} > 5; 
  \]

- Recall that we assume independence of terms and so the filter’s RF is the product of the terms’ RFs

- Discuss with a neighbor:
  - How many histograms would we need?
  - Suppose you want to create new histogram(s) on an existing relation. Brainstorm what you would need to do. Think of the functionality from Exercise 2!

Lab 3: SimpleDb Optimizer

- Statistics, like the Catalog, are memory-only in SimpleDb

- Generated when Parser initialized

- You can assume a fixed number of buckets

Exercise 1: understand formation of physical query plan in SimpleDb

Exercise 2-3: histogram for statistics, and setting up TableStats
(Remaining slides shifted to next lecture)