Selection

Recall: Notions of Clustering

Clustered-file organization

- tuples of one relation R are stored in blocks together with tuples of some other relation S with which they share a common value
  - to optimize the join of the two relations

Clustered relation (= contiguous storage)

- tuples of the relation are stored in blocks that are exclusively or at least predominantly devoted to storing that relation

Clustering index

- an index in which the tuples having a given value of the search key appear in blocks that are largely devoted to storing tuples with that search-key value

Selection (1)

Key decision: shall we use an index and when we have the choice which one?

Task: Implementation of $\sigma_C(R)$, Metric: Disk I/Os

Options:

- Scan the complete relation
  - B(R) if R is clustered
  - T(R) if R is not clustered
- Condition C is an equality term such as a = 10, the a-value we search is uniformly distributed over V(R, a) and we use an index on attribute a
  - B(R) / V(R, a) if the index is clustering
  - T(R) / V(R, a) if the index is not clustering
- Condition C is an inequality term such as b < 20 and we use an index on attribute b
  - B(R) / 3 if the index is clustering
  - T(R) / 3 if the index is not clustering

Selection (2)

In case of index usage, we also have to account for disk I/O's to read some index blocks.

If we have several indexes available chose the one that produces the better result

- load tuples based on this index and check other conditions in main memory

Selection: Example (1)

Selection: $\sigma_{X=1 \text{ AND } Y=2 \text{ AND } Z<5}$ (R)

T(R) = 5000

B(R) = 200

V(R, x) = 100

V(R, y) = 500

R is clustered

There are indexes on each of x, y, and z but only the index of z is clustering

Selection: Example (2)

Scan relation: cost is B(R) = 200 I/Os

Use index on x to find those tuples with x=1 and check for each tuple the rest of the condition:
cost is T(R) / V(R, x) = 50 I/Os

Use index on y to find those tuples with y=2 and check for each tuple the rest of the condition:
cost is T(R) / V(R, y) = 10 I/Os

Use index on z to find those tuples with z<5 and check for each tuple the rest of the condition:
cost is B(R) / 3 = 66 + 2/3 I/Os