22 – Cost Model

1. Cost Models
    A. Generate an estimate of the cost of the query plans
    B. Types
        i. Physical cost (CPU cycles, IO, cache misses, etc…)
        ii. Logical cost (result size)
        iii. Algorithmic cost(complexity)
    C. Disk access dominate the cost of disk-based DBMS
    D. In-Memory DB have no IO costs, but must consider CPU and memory access costs
    E. Commercial DB's cost model
        i. Postgres (Disk)
        combinations of CPU and IO cost with magic constant factors
        ii. IBM DB2 (Disk)
        DB characteristics in system catalogs
        hardware environment, storage device characteristics
        communication bandwidth, memory resources,
        concurrency environments
        iii. Smallbase (In-Memory)
        two phase model (identify execution primitives, microbenchmarks)
2. Cost Estimation
    A. Selectivity
    rate of data accessed for a predicate
    can be estimated by domain constraints, precomputed statistics, etc
        i. Approximations
        maintaining exact statistics is expensive
        Using approximate data is enough to estimate cost

ii. Sampling
run query on subset of data set to estimate selectivity
can use read only copy or real sample table

B. Result cardinality
The number of tuples that will be generated per operator

i. Uniform data

ii. Independent predicate (problem on correlated attributes)

iii. Inclusion principle

iv. Column group statistics
can track statistics for groups of attributes together
(supported in commercial systems)

C. Estimation problem

i. If some of operator's cost estimations go wrong, all estimation will be affected by this gap

ii. Estimation problem can cause execution slowdown

iii. Optimizer is more important than a fast engine.
sequential scan and hash joins are robust execution model