Relational Query Optimization

CS186
R & G Chapters 12/15

Review

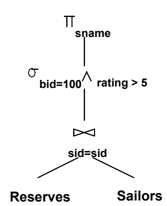
- **Implementation of single Relational Operations**
- **Choices depend on indexes, memory, stats,…**
- **Joins**
  - Blocked nested loops:
    - simple, exploits extra memory
  - Indexed nested loops:
    - best if 1 rel small and one indexed
  - Sort/Merge Join
    - good with small amount of memory, bad with duplicates
  - Hash Join
    - fast (enough memory), bad with skewed data

Query Optimization Overview

- **Query can be converted to relational algebra**
- **Rel. Algebra converts to tree, joins form branches**
- **Each operator has implementation choices**
- **Operators can also be applied in different order!**

```
SELECT S.sname
FROM Reserves R, Sailors S
WHERE R.sid=S.sid AND
   R.bid=100 AND S.rating>5
```

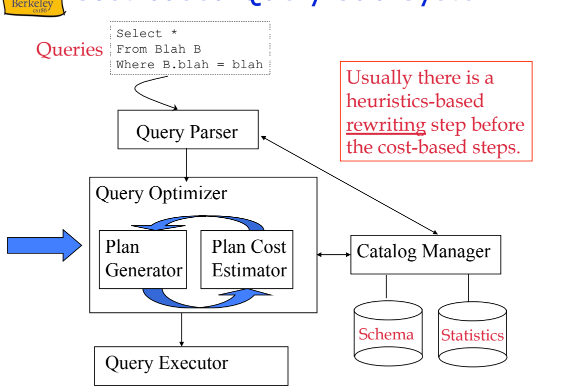
$$\pi_{(sname)}\sigma_{(bid=100 \wedge rating > 5)}(\text{Reserves} \bowtie \text{Sailors})$$



Query Optimization Overview (cont.)

- ***Plan*: *Tree of R.A. ops (and some others) with choice of algorithm for each op.***
  - Recall: Iterator interface (next()!)
- **Three main issues:**
  - For a given query, what plans are considered?
  - How is the cost of a plan estimated?
  - How do we "search" in the "plan space"?
- **Ideally: Want to find best plan.**
- **Reality: Avoid worst plans!**

Cost-based Query Sub-System



Let's go through some examples

- **Just to get a flavor…**

Schema for Examples

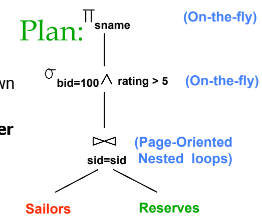
Sailors (<u>*sid*: integer</u>, *sname*: string, *rating*: integer, *age*: real)
Reserves (<u>*sid*: integer, *bid*: integer, *day*: dates</u>, *rname*: string)

- **As seen in previous lectures…**
- **Reserves:**
  - Each tuple is 40 bytes long, 100 tuples per page, 1000 pages.
  - Assume there are 100 boats
- **Sailors:**
  - Each tuple is 50 bytes long, 80 tuples per page, 500 pages.
  - Assume there are 10 different ratings
- **Assume we have 5 pages in our buffer pool!**

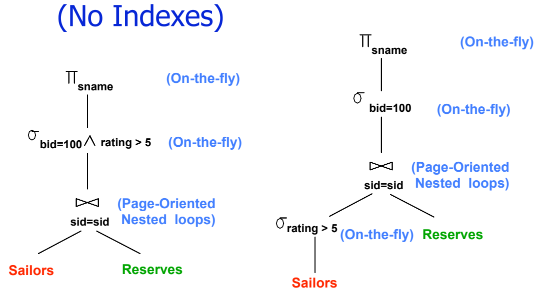
Motivating Example

```
SELECT S.sname
FROM Reserves R, Sailors S
WHERE R.sid=S.sid AND
   R.bid=100 AND S.rating>5
```

- **Cost: 500+500*1000 I/Os**
- **By no means the worst plan!**
- **Misses several opportunities:**
  - selections could be 'pushed' down
  - no use made of indexes
- ***Goal of optimization:* Find faster plans that compute the same answer.**

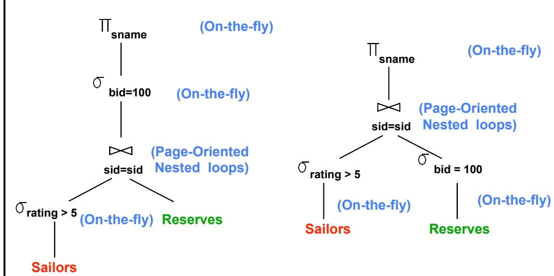


Alternative Plans – Push Selects (No Indexes)



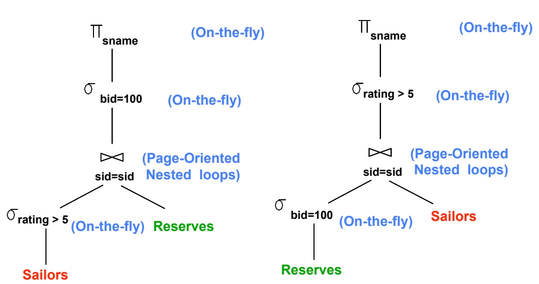
500,500 IOs   250,500 IOs

Alternative Plans – Push Selects (No Indexes)



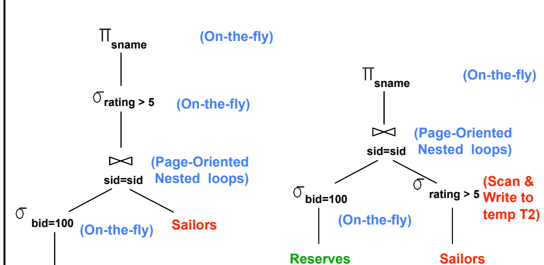
250,500 IOs   250,500 IOs

Alternative Plans – Push Selects (No Indexes)



250,500 IOs   6000 IOs

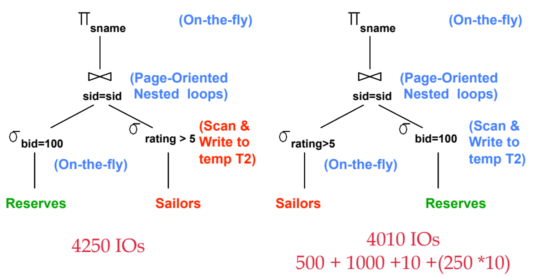
Alternative Plans – Push Selects (No Indexes)



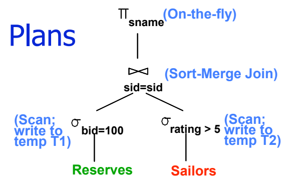
6000 IOs

4250 IOs
1000 + 500+ 250 + (10 * 250)

Alternative Plans – Push Selects (No Indexes)

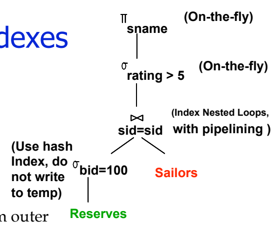


More Alternative Plans (No Indexes)



- ***Sort Merge Join***
- **With 5 buffers, cost of plan:**
  - Scan Reserves (1000) + write temp T1 (10 pages) = 1010.
  - Scan Sailors (500) + write temp T2 (250 pages) = 750.
  - Sort T1 (2*2*10) + sort T2 (2*4*250) + merge (10+250) = 2300
  - Total: 4060 page I/Os.
- **If use BNL join, join = 10+4*250, total cost = 2770.**
- **Can also `push' projections, but must be careful!**
  - T1 has only *sid*, T2 only *sid*, *sname*:
  - T1 fits in 3 pgs, cost of BNL under 250 pgs, total < 2000.

More Alt Plans: Indexes



- **With clustered index on *bid* of Reserves, we get:**
  - 100,000/100 = 1000 tuples on 1000/100 = 10 pages.
- **INL with outer not materialized.**
  - Projecting out unnecessary fields from outer doesn't help.

- Join column *sid* is a key for Sailors.
  - At most one matching tuple, unclustered index on *sid* OK.
- Decision not to push *rating>5* before the join is based on availability of *sid* index on Sailors.
- Cost: Selection of Reserves tuples (10 I/Os); then, for each, must get matching Sailors tuple (1000*1.2); total 1210 I/Os.

Summing up

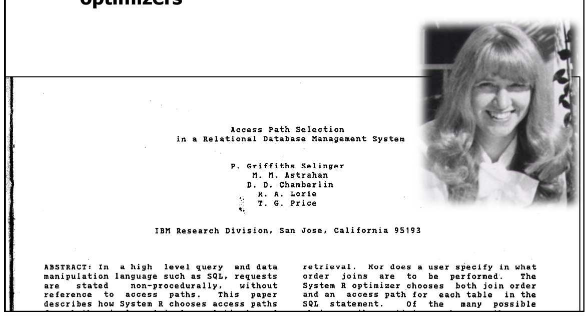
- **There are *lots* of plans**
  - Even for a relatively simple query
- **People tend to think they can pick good ones by hand**
  - MapReduce is based on that assumption
- **Not so clear that's true!**
  - Machines are better at enumerating options than people
  - But we will see soon how optimizers make simplifying assumptions

What is needed for optimization?

- **A closed set of operators**
  - Relational ops (table in, table out)
  - Encapsulation (e.g. based on iterators)
- **Plan space**
  - Based on relational equivalences, different implementations
- **Cost Estimation, based on**
  - Cost formulas
  - Size estimation, in turn based on
    - Catalog information on base tables
    - Selectivity (Reduction Factor) estimation
- **A search algorithm**
  - To sift through the plan space and find lowest cost option!

Query Optimization

- **Will focus on "System R" (Selinger) style optimizers**



Access Path Selection
in a Relational Database Management System

P. Griffiths Selinger
M. M. Astrahan
D. D. Chamberlin
R. A. Lorie
T. G. Price

IBM Research Division, San Jose, California 95193

ABSTRACT: In a high level query and data manipulation language such as SQL, requests are stated non-procedurally, without reference to access paths. This paper describes how System R chooses access paths

retrieval. Nor does a user specify in what order joins are to be performed. The System R optimizer chooses both join order and an access path for each table in the SQL statement. Of the many possible

Highlights of System R Optimizer

- **Impact:**
  - Most widely used currently; works well for 10-15 joins.
- **Cost estimation:**
  - Very inexact, but works ok in practice.
  - Statistics in system catalogs used to estimate cost of operations and result sizes.
  - Considers combination of CPU and I/O costs.
  - System R's scheme has been improved since that time.
- **Plan Space: Too large, must be pruned.**
  - Many plans share common, "overpriced" subtrees
    - ignore them all!
  - In some implementations, only the space of *left-deep plans* is considered.
  - Cartesian products avoided in some implementations.

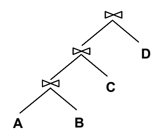
Query Blocks: Units of Optimization

- Break query into *query blocks*
- Optimized one block at a time
- Uncorrelated nested blocks computed once
- Correlated nested blocks like function calls
  - But sometimes can be "decorrelated"
  - Beyond the scope of CS186!

```
SELECT S.sname
FROM Sailors S
WHERE S.age IN
    (SELECT MAX (S2.age)
     FROM Sailors S2
     GROUP BY S2.rating)
```

Outer block   Nested block

- For each block, the plans considered are:
  - All available access methods, for each relation in FROM clause.
  - All *left-deep join trees*
    - right branch always a base table
    - consider all join orders and join methods



Schema for Examples

Sailors (<u>*sid*: integer</u>, *sname*: string, *rating*: integer, *age*: real)
Reserves (<u>*sid*: integer, *bid*: integer, *day*: dates</u>, *rname*: string)

- **Reserves:**
  - Each tuple is 40 bytes long, 100 tuples per page, 1000 pages. 100 distinct bids.
- **Sailors:**
  - Each tuple is 50 bytes long, 80 tuples per page, 500 pages. 10 ratings, 40,000 sids.

Translating SQL to Relational Algebra

```
SELECT S.sid, MIN (R.day)
FROM Sailors S, Reserves R, Boats B
WHERE S.sid = R.sid AND R.bid = B.bid AND B.color = "red"
GROUP BY S.sid

HAVING COUNT (*) >= 2
```

For each sailor with at least two reservations for red boats, find the sailor id and the earliest date on which the sailor has a reservation for a red boat.

Translating SQL to Relational Algebra

```
SELECT S.sid, MIN (R.day)
FROM Sailors S, Reserves R, Boats B
WHERE S.sid = R.sid AND R.bid = B.bid AND B.color = "red"
GROUP BY S.sid

HAVING COUNT (*) >= 2
```

$$\pi_{\text{S.sid, MIN(R.day)}}(\text{HAVING}_{\text{COUNT(*)>2}}(\text{GROUP BY}_{\text{S.Sid}}(\sigma_{\text{B.color = "red"}}(\text{Sailors} \bowtie \text{Reserves} \bowtie \text{Boats}))))$$

Relational Algebra Equivalences

- **Allow us to choose different join orders and to `push' selections and projections ahead of joins.**
- **Selections:**
  - $\sigma_{c1\wedge\ldots\wedge cn}(R) \equiv \sigma_{c1}(\ldots(\sigma_{cn}(R))\ldots)$ (*cascade*)
  - $\sigma_{c1}(\sigma_{c2}(R)) \equiv \sigma_{c1}(\sigma_{c1}(R))$ (*commute*)
- **Projections:**
  - $\pi_{a1}(R) \equiv \pi_{a1}(\ldots(\pi_{a1, \ldots, an}(R))\ldots)$ (*cascade*)
- **Cartesian Product**
  - $R \times (S \times T) \equiv (R \times S) \times T$ (associative)
  - $R \times S \equiv S \times R$ (commutative)
  - *This means we can do joins in any order.*
    - But...beware of cartesian product!

More Equivalences

- **Eager projection**
  - Can cascade and "push" some projections thru selection
  - Can cascade and "push" some projections below one side of a join
  - Rule of thumb: can project anything not needed "downstream"
- **Selection on a cross-product is equivalent to a join.**
  - If selection is comparing attributes from each side
- **A selection on attributes of R commutes with $R \bowtie S$.**
  - i.e., $\sigma(R \bowtie S) \equiv \sigma(R) \bowtie S$
  - but only if the selection doesn't refer to S!

Cost Estimation

- **For each plan considered, must estimate total cost:**
  - Must estimate *cost* of each operation in plan tree.
    - Depends on input cardinalities.
    - We've already discussed this for various operators
      - sequential scan, index scan, joins, etc.
  - Must estimate *size of result* for each operation in tree!
    - Use information about the input relations.
    - For selections and joins, assume independence of predicates.
  - In System R, cost is boiled down to a single number consisting of #I/O + *factor* * #CPU instructions
- **Q: Is "cost" the same as estimated "run time"?**

Statistics and Catalogs

- **Need info on relations and indexes involved.**
- *Catalogs* **typically contain at least:**
  - # tuples (**NTuples**) and # pages (**NPages**) per rel'n.
  - # distinct key values (**NKeys**) for each index.
  - low/high key values (**Low/High**) for each index.
  - Index height (**IHeight**) for each tree index.
  - # index pages (**INPages**) for each index.
- **Catalogs updated periodically.**
  - Too expensive to do continuously
  - Lots of approximation anyway, so a little slop here is ok.
- **Modern systems do more**
  - Estimate these quantities in absence of indexes
  - Keep more detailed information on data values
    - e.g., histograms

Size Estimation and Reduction Factors

```
SELECT attribute list
FROM relation list
WHERE term1 AND ... AND termk
```

- **Consider a query block:**
- **Maximum # tuples in result is the product of the cardinalities of relations in the FROM clause.**
- ***Reduction factor (RF)* associated with each *term* reflects the impact of the *term* in reducing result size. *Result cardinality* = Max # tuples * product of all RF's.**
- ***RF usually called "selectivity"***
  - only R&G seem to call it Reduction Factor
  - beware of confusion between "high selectivity" as defined here and "highly selective" in common English!

Result Size Estimation

- *Result cardinality* = **Max # tuples * product of all RF's.**
- Term *col=value* (given index I on *col*)
  RF = *1/NKeys(I)*
- Term *col1=col2* (handy for joins too…)
  RF = *1/MAX(NKeys(I1), NKeys(I2))*
- Term *col>value*
  RF = *(High(I)-value)/(High(I)-Low(I))*

  Implicit assumptions: values are uniformly distributed and *terms* are independent!
- *Note, if missing indexes, assume* ***1/10****!!!*

Postgres 8: include/utils/selfuncs.h

```
/* default selectivity estimate
   for equalities such as "A = b"
   */
#define DEFAULT_EQ_SEL  0.005

/* default selectivity estimate
   for inequalities such as "A <
   b" */
#define DEFAULT_INEQ_SEL
   0.3333333333333333

/* default selectivity estimate
   for range inequalities "A > b
   AND A < c" */
#define DEFAULT_RANGE_INEQ_SEL
   0.005

/* default selectivity estimate
   for pattern-match operators
   such as LIKE */
#define DEFAULT_MATCH_SEL   0.005

/* default number of distinct
   values in a table */
#define DEFAULT_NUM_DISTINCT  200

/* default selectivity estimate
   for boolean and null test
   nodes */
#define DEFAULT_UNK_SEL
   0.005
#define DEFAULT_NOT_UNK_SEL
   (1.0 - DEFAULT_UNK_SEL)
```

Backend/optimizer/path/clausesel.c

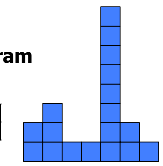
```
/*
 *
 * THIS IS A HACK TO GET V4 OUT THE DOOR.
 *        -- JMH 7/9/92
 */
s1 = (Selectivity) 0.3333333;
```

Reduction Factors & Histograms

- **For better estimation, use a histogram**

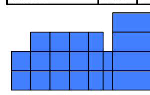
equiwidth

No. of Values	2	3	3	1	8	2	1
Value	0-.99	1-1.99	2-2.99	3-3.99	4-4.99	5-5.99	6-6.99



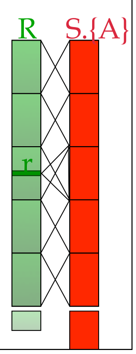
equidepth

No. of Values	2	3	3	3	3	2	4
Value	0-.99	1-1.99	2-2.99	3-4.05	4.06-4.67	4.68-4.99	5-6.99



Think through estimation for joins

- Term *col1=col2*
  - RF = *1/MAX(NKeys(I1), NKeys(I2))*
- **Q: Given a join of R and S, what is the range of possible result sizes (in #of tuples)?**
  - If join is on a key for R (and a Foreign Key in S)?
    - A common case, can treat it specially
- **General case: join on {A} ({A} is key for neither)**
  - estimate each tuple r of R generates NTuples(S)/ NKeys(A,S) result tuples, so…
    NTuples(**R**) * NTuples(**S**)/NKeys(A,**S**)
  - but can also consider it starting with S, yielding:
    NTuples(**S**) * NTuples(**R**)/NKeys(A,**R**)
  - If these two estimates differ, take the lower one!
    - Q: Why?



Enumeration of Alternative Plans

- **There are two main cases:**
  - Single-relation plans (base case)
  - Multiple-relation plans (induction)
- **Single-table queries include selects, projects, and grouping/aggregate ops:**
  - Consider each available access path (file scan / index)
    - Choose the one with the least estimated cost
  - Selection/Projection done on the fly
  - Result pipelined into grouping/aggregation

Cost Estimates for Single-Relation Plans

- **Index I on primary key matches selection:**
  - *Cost is Height(I)+1 for a B+ tree.*
- **Clustered index I matching one or more selects:**
  - *(NPages(I)+NPages(R)) * product of RF's of matching selects.*
- **Non-clustered index I matching one or more selects:**
  - *(NPages(I)+NTuples(R)) * product of RF's of matching selects.*
- **Sequential scan of file:**
  - *NPages(R).*

☞ *Recall: Must also charge for duplicate elimination if required*

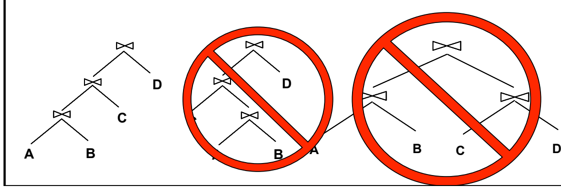
Example

```
SELECT S.sid
FROM Sailors S
WHERE
S.rating=8
```

- **If we have an index on *rating*:**
  - Cardinality = (1/NKeys(I)) * NTuples(R) = (1/10) * 40000 tuples
  - Clustered index: (1/NKeys(I)) * (NPages(I)+NPages(R)) = (1/10) * (50+500) = 55 pages are retrieved. (This is the ***cost***.)
  - Unclustered index: (1/NKeys(I)) * (NPages(I)+NTuples(R)) = (1/10) * (50+40000) = 401 pages are retrieved.
- **If we have an index on *sid*:**
  - Would have to retrieve all tuples/pages. With a clustered index, the cost is 50+500, with unclustered index, 50+40000.
- **Doing a file scan:**
  - We retrieve all file pages (500).

Queries Over Multiple Relations

- **A heuristic decision in System R: *only left-deep join trees* are considered.**
  - As the number of joins increases, the number of alternative plans grows rapidly; *we need to restrict the search space.*
  - Left-deep trees allow us to generate all *fully pipelined* plans.
    - Intermediate results not written to temporary files.
    - Not all left-deep trees are fully pipelined (e.g., SM join).



Enumeration of Left-Deep Plans

- **Left-deep plans differ in**
  - the order of relations
  - the access method for each relation
  - the join method for each join.
- **Enumerated using N passes (if N relations joined):**
  - Pass 1: Find best 1-relation plan for each relation.
  - Pass i: Find best way to join result of an (i -1)-relation plan (as outer) to the i'th relation. (i between 2 and N.)
- **For each subset of relations, retain only:**
  - Cheapest plan overall, plus
  - Cheapest plan for each *interesting order* of the tuples.

The Dynamic Programming Table

Subset of tables in FROM clause	Interesting-order columns	Best plan	Cost
{R, S}	<none>	hashjoin(R,S)	1000
{R, S}	<R.a, S.b>	sortmerge(R,S)	1500

A Note on "Interesting Orders"

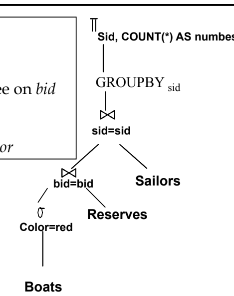
- **An intermediate result has an "interesting order" if it is sorted by any of:**
  - ORDER BY attributes
  - GROUP BY attributes
  - Join attributes of *yet-to-be-added* (downstream) joins

Enumeration of Plans (Contd.)

- **Match an *i* -1 way plan with another table *only if***
  a) there is a join condition between them, *or*
  b) all predicates in WHERE have been used up.
    - i.e., avoid Cartesian products if possible.
- **ORDER BY, GROUP BY, aggregates etc. handled as a final step**
  - via `interestingly ordered' plan if chosen (free!)
  - or via an additional sort/hash operator
- **In spite of pruning plan space, this approach is still exponential in the # of tables.**
- **Recall that in practice, COST considered is *#IOs + factor * CPU Inst***

Example

Sailors:
Hash, B+ on *sid*
Reserves:
Clustered B+ tree on *bid*
B+ on *sid*
Boats
B+, Hash on *color*



Select S.sid, COUNT(*) AS number
FROM Sailors S, Reserves R, Boats B
WHERE S.sid = R.sid AND R.bid = B.bid
AND B.color = "red"
GROUP BY S.sid

- **Pass1: Best plan(s) for accessing each relation**
  - Reserves, Sailors: File Scan
  - Q: What about Clustered B+ on Reserves.bid???
  - Boats: B+ tree & Hash on color

Pass 1

- **Find best plan for each relation in isolation:**
  - Reserves, Sailors: File Scan
  - Boats: B+ tree & Hash on color

Pass 2

- **For each plan in pass 1, generate plans joining another relation as the inner, using all join methods (and matching inner access methods)**
  - File Scan Reserves (outer) with Boats (inner)
  - File Scan Reserves (outer) with Sailors (inner)
  - File Scan Sailors (outer) with Boats (inner)
  - File Scan Sailors (outer) with Reserves (inner)
  - Boats hash on color with Sailors (inner)
  - Boats Btree on color with Sailors (inner)
  - Boats hash on color with Reserves (inner) (sort-merge)
  - Boats Btree on color with Reserves (inner) (BNL)
- **Retain cheapest plan for each (pair of relations, order)**

Pass 3 and beyond

- **Using Pass 2 plans as outer relations, generate plans for the next join**
  - E.g. Boats hash on color with Reserves (bid) (inner) (sortmerge))
    inner Sailors (B-tree sid) sort-merge
- **Then, add cost for groupby/aggregate:**
  - This is the cost to sort the result by sid, *unless it has already been sorted by a previous operator.*
- **Then, choose the cheapest plan**

Points to Remember

- **Want to understand DB design (tables, indexes)?**
  - *Must* understand query optimization
- **Two parts to optimizing a query:**
  - Consider a set of alternative plans, pruning search
    - E.g., left-deep plans only
    - avoid Cartesian products.
    - Prune plans with *interesting orders* separate from unordered plans
  - Must estimate cost of each plan that is considered.
    - Output cardinality and cost for each plan node.
    - *Key issues*: Statistics, indexes, operator implementations.

Points to Remember

- **Single-relation queries:**
  - All access paths considered, cheapest is chosen.
  - *Issues*:
    - Selections that *match* index
    - whether index key has all needed fields
    - whether index provides tuples in an interesting order.

More Points to Remember

- **Multiple-relation queries:**
  - All single-relation plans are first enumerated.
    - Selections/projections considered as early as possible.
  - Use best 1-way plans to form 2-way plans. Prune losers.
  - Use best (i-1)-way plans and best 1-way plans to form i-way plans
  - At each level, for each subset of relations, retain:
    - best plan for each interesting order (including no order)

Summary

- **Optimization is the reason for the lasting power of the relational system**
- **But it is primitive in some ways**
- **New areas: Smarter summary statistics (fancy histograms and "sketches"), auto-tuning statistics, adaptive runtime re-optimization (e.g. *eddies*), multi-query optimization, parallel scheduling issues, etc.**