Introduction to Database Systems CSE 444

Lecture 22: Query Optimization

November 26-30, 2007

Outline

- An example
- Query optimization: algebraic laws 16.2
- Cost-based optimization 16.5, 16.6
- Cost estimation: 16.4

Example

Product(<u>pname</u>, maker), **Company**(<u>cname</u>, city)

```
Select Product.pname
From  Product, Company
Where Product.maker=Company.cname
   and  Company.city = “Seattle”
```

- How do we execute this query ?

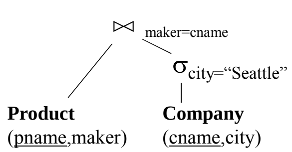
Example

Product(<u>pname</u>, maker), **Company**(<u>cname</u>, city)

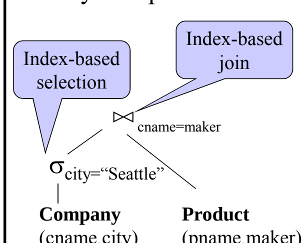
Assume:

Clustered index: **Product**.<u>pname</u>, **Company**.<u>cname</u>
Unclustered index: **Product**.maker, **Company**.city

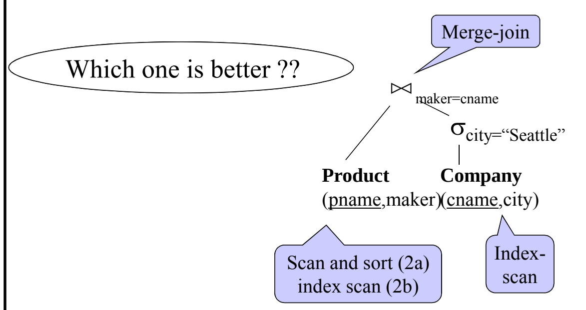
Logical Plan:

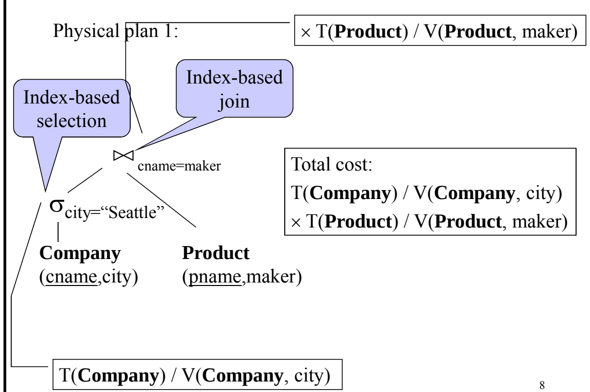


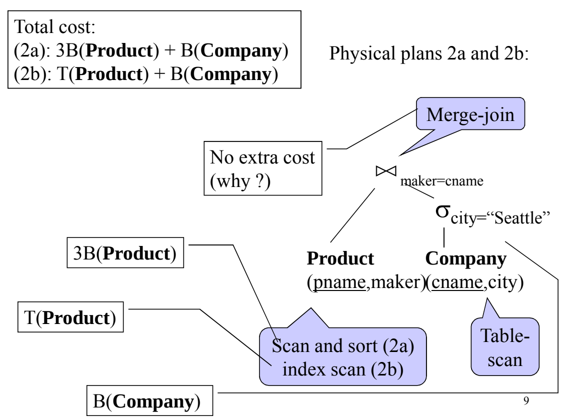
Physical plan 1:



Physical plans 2a and 2b:







Plan 1: T(**Company**)/V(**Company**,city) × T(**Product**)/V(**Product**,maker)
Plan 2a: B(**Company**) + 3B(**Product**)
Plan 2b: B(**Company**) + T(**Product**)

Which one is better ??

It depends on the data !!

Example

T(**Company**) = 5,000 B(**Company**) = 500 M = 100
T(**Product**) = 100,000 B(**Product**) = 1,000

We may assume V(**Product**, maker) ≈ T(**Company**) (why ?)

- Case 1: V(**Company**, city) ≈ T(**Company**)

V(**Company**,city) = 2,000

- Case 2: V(**Company**, city) << T(**Company**)

V(**Company**,city) = 20

Which Plan is Best ?

Plan 1: T(**Company**)/V(**Company**,city) × T(**Product**)/V(**Product**,maker)
Plan 2a: B(**Company**) + 3B(**Product**)
Plan 2b: B(**Company**) + T(**Product**)

Case 1:

Case 2:

Lessons

- Need to consider several physical plan
  - even for one, simple logical plan
- No magic "best" plan: depends on the data
- In order to make the right choice
  - need to have ***statistics*** over the data
  - the B's, the T's, the V's

13

Query Optimzation

- Have a SQL query Q
- Create a plan P
- Find equivalent plans P = P' = P'' = …
- Choose the "cheapest".

HOW ??

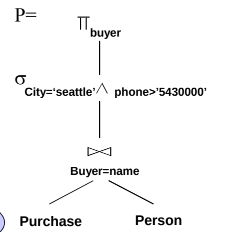
14

Logical Query Plan

```
SELECT  P.buyer
FROM Purchase P, Person Q
WHERE P.buyer=Q.name AND
      P.city='seattle' AND
      Q.phone > '5430000'
```

Purchasse(buyer, city)
Person(name, phone)

P=



In class:
find a "better" plan P'

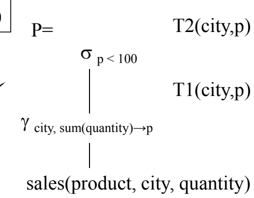
15

Logical Query Plan

Q=

```
SELECT city, sum(quantity)
FROM sales
GROUP BY city
HAVING sum(quantity) < 100
```

P=



In class:
find a "better" plan P'

16

The three components of an optimizer

We need three things in an optimizer:

- Algebraic laws
- An optimization algorithm
- A cost estimator

17

Algebraic Laws (incomplete list)

- Commutative and Associative Laws

  $R \cup S = S \cup R, \;\; R \cup (S \cup T) = (R \cup S) \cup T$

  $R \;|\times|\; S = S \;|\times|\; R, \;\; R \;|\times|\; (S \;|\times|\; T) = (R \;|\times|\; S) \;|\times|\; T$

- Distributive Laws

  $R \;|\times|\; (S \cup T) \;=\; (R \;|\times|\; S) \cup (R \;|\times|\; T)$

18

Algebraic Laws (incomplete list)

- Laws involving selection:
  - $\sigma_{C \text{ AND } C'}(R) = \sigma_C(\sigma_{C'}(R))$
  - $\sigma_{C \text{ OR } C'}(R) = \sigma_C(R) \cup \sigma_{C'}(R)$
- When C involves only attributes of R
  - $\sigma_C(R \mid\times\mid S) = \sigma_C(R) \mid\times\mid S$
  - $\sigma_C(R - S) = \sigma_C(R) - S$
  - $\sigma_C(R \mid\times\mid S) = \sigma_C(R) \mid\times\mid S$

19

Algebraic Laws

- Example: R(A, B, C, D), S(E, F, G)
  - $\sigma_{F=3}(R \mid\times\mid_{D=E} S) =$?
  - $\sigma_{A=5 \text{ AND } G=9}(R \mid\times\mid_{D=E} S) =$?

20

Algebraic Laws

- Laws involving projections
  - $\Pi_M(R \mid\times\mid S) = \Pi_M(\Pi_P(R) \mid\times\mid \Pi_Q(S))$
  - $\Pi_M(\Pi_N(R)) = \Pi_{M,N}(R)$
- Example R(A,B,C,D), S(E, F, G)
  - $\Pi_{A,B,G}(R \mid\times\mid_{D=E} S) = \Pi_{?}(\Pi_?(R) \mid\times\mid_{D=E} \Pi_?(S))$

21

Algebraic Laws

- Laws involving grouping and aggregation:
  - $\delta(\gamma_{A,\, agg(B)}(R)) = \gamma_{A,\, agg(B)}(R)$
  - $\gamma_{A,\, agg(B)}(\delta(R)) = \gamma_{A,\, agg(B)}(R)$ if agg is "duplicate insensitive"
- Which of the following are "duplicate insensitive" ? sum, count, avg, min, max

$$\gamma_{A,\, agg(D)}(R(A,B) \mid\times\mid_{B=C} S(C,D)) = \gamma_{A,\, agg(D)}(R(A,B) \mid\times\mid_{B=C} (\gamma_{C,\, agg(D)}S(C,D)))$$

22

Cost-based Optimizations

- Main idea: apply algebraic laws, until estimated cost is minimal
- Practically: start from partial plans, introduce operators one by one
  - Will see in a few slides
- Problem: there are too many ways to apply the laws, hence too many (partial) plans

23

Cost-based Optimizations

Approaches:

- **Top-down**: the partial plan is a top fragment of the logical plan
- **Bottom up**: the partial plan is a bottom fragment of the logical plan

24

Dynamic Programming

Originally proposed in System R (the first research prototype for a relational database system -- late 70s)

- Only handles single block queries:

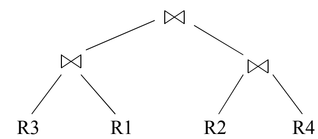
```
SELECT list
FROM    list
WHERE cond1 AND cond2 AND . . . AND condk
```

- Heuristics: selections down, projections up
- Dynamic programming: ***join reordering***

25

Join Trees

- R1 |×| R2 |×| …. |×| Rn
- Join tree:

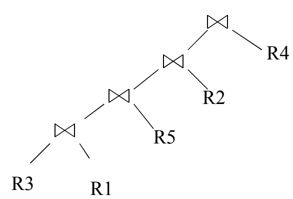


- A plan = a join tree
- A partial plan = a subtree of a join tree

26

Types of Join Trees

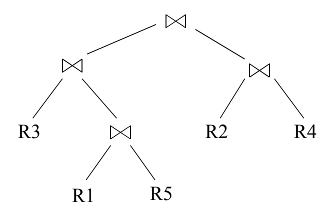
- Left deep:



27

Types of Join Trees

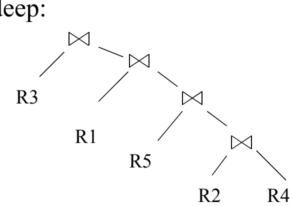
- Bushy:



28

Types of Join Trees

- Right deep:



29

Dynamic Programming

- Given: a query R1 |×| R2 |×| … |×| Rn
- Assume we have a function cost() that gives us the cost of every join tree
- Find the best join tree for the query

30

Dynamic Programming

- Idea: for each subset of {R1, …, Rn}, compute the best plan for that subset
- In increasing order of set cardinality:
  - Step 1: for {R1}, {R2}, …, {Rn}
  - Step 2: for {R1,R2}, {R1,R3}, …, {Rn-1, Rn}
  - …
  - Step n: for {R1, …, Rn}
- It is a bottom-up strategy
- A subset of {R1, …, Rn} is also called a *subquery*

31

Dynamic Programming

- For each subquery Q $\subseteq$ {R1, …, Rn} compute the following:
  - Size(Q)
  - A best plan for Q: Plan(Q)
  - The cost of that plan: Cost(Q)

32

Dynamic Programming

- **Step 1**: For each $\{R_i\}$ do:
  - Size($\{R_i\}$) = B(R_i)
  - Plan($\{R_i\}$) = R_i
  - Cost($\{R_i\}$) = (cost of scanning R_i)

33

Dynamic Programming

- **Step i**: For each Q $\subseteq \{R_1, \ldots, R_n\}$ of cardinality i do:
  - Compute Size(Q) (later…)
  - For every pair of subqueries Q', Q'' s.t. Q = Q' $\cup$ Q''
    compute cost(Plan(Q') |×| Plan(Q''))
  - Cost(Q) = the smallest such cost
  - Plan(Q) = the corresponding plan

34

Dynamic Programming

- Return Plan($\{R_1, \ldots, R_n\}$)

35

Dynamic Programming

To illustrate, we will make the following simplifications:

- Cost(P_1 |×| P_2) = Cost(P_1) + Cost(P_2) + size(intermediate result(s))
- Intermediate results:
  - If P_1 = a join, then the size of the intermediate result is size(P_1), otherwise the size is 0
  - Similarly for P_2
- Cost of a scan = 0

36

Dynamic Programming

- Example:
- Cost(R5 |×| R7) = 0 (no intermediate results)
- Cost((R2 |×| R1) |×| R7)
  = Cost(R2 |×| R1) + Cost(R7) + size(R2 |×| R1)
  = size(R2 |×| R1)

Dynamic Programming

- Relations: R, S, T, U
- Number of tuples: 2000, 5000, 3000, 1000
- Size estimation: T(A |×| B) = 0.01*T(A)*T(B)

Subquery	Size	Cost	Plan
RS			
RT			
RU			
ST			
SU			
TU			
RST			
RSU			
RTU			
STU			
RSTU			

Subquery	Size	Cost	Plan
RS	100k	0	RS
RT	60k	0	RT
RU	20k	0	RU
ST	150k	0	ST
SU	50k	0	SU
TU	30k	0	TU
RST	3M	60k	(RT)S
RSU	1M	20k	(RU)S
RTU	0.6M	20k	(RU)T
STU	1.5M	30k	(TU)S
RSTU	30M	60k+50k=110k	(RT)(SU)

Reducing the Search Space

- Left-linear trees v.s. Bushy trees
- Trees without cartesian product

Example: R(A,B) |×| S(B,C) |×| T(C,D)

Plan: (R(A,B) |×| T(C,D)) |×| S(B,C) has a cartesian product – most query optimizers will not consider it

Dynamic Programming: Summary

- Handles only join queries:
  - Selections are pushed down (i.e. early)
  - Projections are pulled up (i.e. late)
- Takes exponential time in general, BUT:
  - Left linear joins may reduce time
  - Non-cartesian products may reduce time further

Rule-Based Optimizers

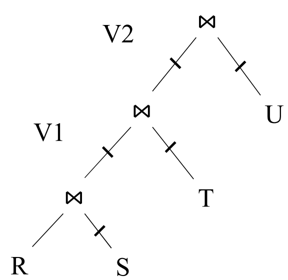
- ***Extensible*** collection of rules
  Rule = Algebraic law with a direction
- Algorithm for firing these rules
  Generate many alternative plans, in some order
  Prune by cost

- Volcano (later SQL Sever)
- Starburst (later DB2)

Completing the Physical Query Plan

- Choose algorithm to implement each operator
  - Need to account for more than cost:
    - How much memory do we have ?
    - Are the input operand(s) sorted ?
- Decide for each intermediate result:
  - To materialize
  - To pipeline

Materialize Intermediate Results Between Operators




```
HashTable ← S
repeat   read(R, x)
         y ← join(HashTable, x)
         write(V1, y)

HashTable ← T
repeat   read(V1, y)
         z ← join(HashTable, y)
         write(V2, z)

HashTable ← U
repeat   read(V2, z)
         u ← join(HashTable, z)
         write(Answer, u)
```

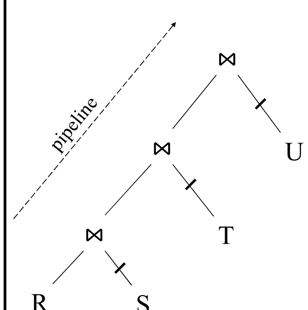

Materialize Intermediate Results Between Operators

Question in class

Given B(R), B(S), B(T), B(U)

- What is the total cost of the plan ?
  - Cost =
- How much main memory do we need ?
  - M =

Pipeline Between Operators




```
HashTable1 ← S
HashTable2 ← T
HashTable3 ← U
repeat   read(R, x)
         y ← join(HashTable1, x)
         z ← join(HashTable2, y)
         u ← join(HashTable3, z)
         write(Answer, u)
```

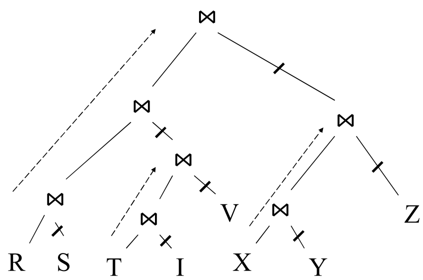

Pipeline Between Operators

Question in class

Given B(R), B(S), B(T), B(U)

- What is the total cost of the plan ?
  - Cost =
- How much main memory do we need ?
  - M =

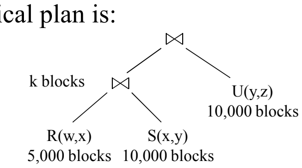
Pipeline in Bushy Trees



49

Example

- Logical plan is:

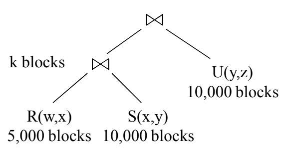


- Main memory M = 101 buffers

50

Example

M = 101



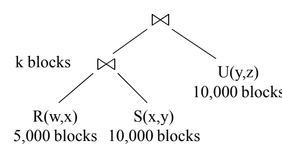
Naïve evaluation:

- 2 partitioned hash-joins
- Cost 3B(R) + 3B(S) + 4k + 3B(U) = 75000 + 4k

51

Example

M = 101



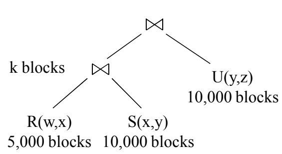
Smarter:

- Step 1: hash R on x into 100 buckets, each of 50 blocks; to disk
- Step 2: hash S on x into 100 buckets; to disk
- Step 3: read each R_i in memory (50 buffer) join with S_i (1 buffer); hash result on y into 50 buckets (50 buffers) -- here we ***pipeline***
- Cost so far: 3B(R) + 3B(S)

52

Example

M = 101



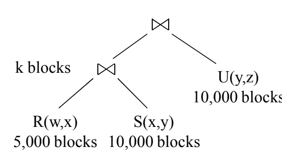
Continuing:

- How large are the 50 buckets on y ? Answer: k/50.
- If k <= 50 then keep all 50 buckets in Step 3 in memory, then:
- Step 4: read U from disk, hash on y and join with memory
- Total cost: 3B(R) + 3B(S) + B(U) = 55,000

53

Example

M = 101



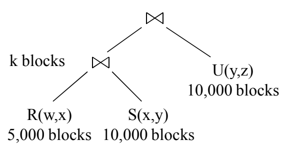
Continuing:

- If 50 < k <= 5000 then send the 50 buckets in Step 3 to disk
  - Each bucket has size k/50 <= 100
- Step 4: partition U into 50 buckets
- Step 5: read each partition and join in memory
- Total cost: 3B(R) + 3B(S) + 2k + 3B(U) = 75,000 + 2k

54

Example

M = 101



Continuing:

- If k > 5000 then materialize instead of pipeline
- 2 partitioned hash-joins
- Cost 3B(R) + 3B(S) + 4k + 3B(U) = 75000 + 4k

Example

Summary:

- If k <= 50, cost = 55,000
- If 50 < k <=5000, cost = 75,000 + 2k
- If k > 5000, cost = 75,000 + 4k

Size Estimation

The problem: Given an expression E, compute T(E) and V(E, A)

- This is hard without computing E
- Will ‘estimate’ them instead

Size Estimation

Estimating the size of a projection

- Easy: $T(\Pi_L(R)) = T(R)$
- This is because a projection doesn’t eliminate duplicates

Size Estimation

Estimating the size of a selection

- $S = \sigma_{A=c}(R)$
  - T(S) san be anything from 0 to T(R) – V(R,A) + 1
  - Estimate: T(S) = T(R)/V(R,A)
  - When V(R,A) is not available, estimate T(S) = T(R)/10
- $S = \sigma_{A<c}(R)$
  - T(S) can be anything from 0 to T(R)
  - Estimate: T(S) = (c - Low(R, A))/(High(R,A) - Low(R,A))T(R)
  - When Low, High unavailable, estimate T(S) = T(R)/3

Size Estimation

Estimating the size of a natural join, $R \mid\times\mid_A S$

- When the set of A values are disjoint, then $T(R \mid\times\mid_A S) = 0$
- When A is a key in S and a foreign key in R, then $T(R \mid\times\mid_A S) = T(R)$
- When A has a unique value, the same in R and S, then $T(R \mid\times\mid_A S) = T(R)\ T(S)$

Size Estimation

Assumptions:

- *Containment of values*: if V(R,A) <= V(S,A), then the set of A values of R is included in the set of A values of S
  - Note: this indeed holds when A is a foreign key in R, and a key in S
- *Preservation of values*: for any other attribute B, V(R $|\times|_A$ S, B) = V(R, B) (or V(S, B))

61

Size Estimation

Assume V(R,A) <= V(S,A)

- Then each tuple t in R joins *some* tuple(s) in S
  - How many ?
  - On average T(S)/V(S,A)
  - t will contribute T(S)/V(S,A) tuples in R $|\times|_A$ S
- Hence T(R $|\times|_A$ S) = T(R) T(S) / V(S,A)

In general: T(R $|\times|_A$ S) = T(R) T(S) / max(V(R,A),V(S,A))

62

Size Estimation

Example:

- T(R) = 10000, T(S) = 20000
- V(R,A) = 100, V(S,A) = 200
- How large is R $|\times|_A$ S ?

Answer: T(R $|\times|_A$ S) = 10000 20000/200 = 1M

63

Size Estimation

Joins on more than one attribute:

- T(R $|\times|_{A,B}$ S) =

  T(R) T(S)/(max(V(R,A),V(S,A))*max(V(R,B),V(S,B)))

64

Histograms

- Statistics on data maintained by the RDBMS
- Makes size estimation much more accurate (hence, cost estimations are more accurate)

65

Histograms

Employee(<u>ssn</u>, name, salary, phone)

- Maintain a histogram on salary:

Salary:	0..20k	20k..40k	40k..60k	60k..80k	80k..100k	> 100k
Tuples	200	800	5000	12000	6500	500

- T(Employee) = 25000, but now we know the distribution

66

Histograms

Ranks(rankName, salary)

- Estimate the size of Employee $|\times|_{Salary}$ Ranks

Employee	0..20k	20k..40k	40k..60k	60k..80k	80k..100k	> 100k
	200	800	5000	12000	6500	500

Ranks	0..20k	20k..40k	40k..60k	60k..80k	80k..100k	> 100k
	8	20	40	80	100	2

Histograms

- Eqwidth

0..20	20..40	40..60	60..80	80..100
2	104	9739	152	3

- Eqdepth

0..44	44..48	48..50	50..56	55..100
2000	2000	2000	2000	2000