Lecture 23: Query Execution

Wednesday, March 8, 2006

Outline

• Query optimization: algebraic laws 16.2
Example

**Product**(pname, maker), **Company**(cname, city)

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

• How do we execute this query?

---

Example

**Product**(pname, maker), **Company**(cname, city)

Assume:

- Clustered index: **Product.pname**, **Company.cname**
- Unclustered index: **Product.maker**, **Company.city**
Logical Plan:

\[ \sigma_{\text{city}=\text{"Seattle"}} \left( \text{Product} \left( \text{pname}, \text{maker} \right) \right) \join \sigma_{\text{maker}=\text{cname}} \left( \text{Company} \left( \text{cname}, \text{city} \right) \right) \]

Physical plan 1:

\[ \sigma_{\text{city}=\text{"Seattle"}} \left( \text{Company} \left( \text{cname}, \text{city} \right) \right) \join \sigma_{\text{cname}=\text{maker}} \left( \text{Product} \left( \text{pname}, \text{maker} \right) \right) \]
Physical plans 2a and 2b:

Which one is better??
Total cost:
(2a): 3B(Product) + B(Company)
(2b): T(Product) + B(Company)

Physical plans 2a and 2b:
- Merge-join
- No extra cost (why?)
- Scan and sort (2a)
- index scan (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

\[
\begin{align*}
T(\text{Company}) &= 5,000 & B(\text{Company}) &= 500 & M &= 100 \\
T(\text{Product}) &= 100,000 & B(\text{Product}) &= 1,000
\end{align*}
\]

We may assume \(V(\text{Product, maker}) \approx T(\text{Company})\) (why ?)

- Case 1: \(V(\text{Company, city}) \approx T(\text{Company})\)
  \[
  V(\text{Company, city}) = 2,000
  \]

- Case 2: \(V(\text{Company, city}) \ll T(\text{Company})\)
  \[
  V(\text{Company, city}) = 20
  \]

Which Plan is Best ?

Plan 1: \(T(\text{Company})/V(\text{Company, city}) \times T(\text{Product})/V(\text{Product, maker})\)
Plan 2a: \(B(\text{Company}) + 3B(\text{Product})\)
Plan 2b: \(B(\text{Company}) + T(\text{Product})\)

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

Query Optimization

• Have a SQL query Q

• Create a plan P

• Find equivalent plans P = P’ = P’’ = …

• Choose the “cheapest”.

HOW ??
Logical Query Plan

\[
\begin{align*}
\text{SELECT} & \quad \text{P.buyer} \\
\text{FROM} & \quad \text{Purchase P, Person Q} \\
\text{WHERE} & \quad \text{P.buyer=Q.name AND} \\
& \quad \text{P.city='seattle' AND} \\
& \quad \text{Q.phone > '5430000'}
\end{align*}
\]

\[P = T_{\text{buyer}} \sigma_{\text{city='seattle' \land phone > '5430000'}} \sigma_{\text{Buyer=\text{name}}} \text{Purchase} \rightarrow \text{Person}\]

In class:
find a “better” plan \(P'\)

Logical Query Plan

\[
\begin{align*}
\text{SELECT} & \quad \text{city, sum(quantity)} \\
\text{FROM} & \quad \text{sales} \\
\text{GROUP BY} & \quad \text{city} \\
\text{HAVING} & \quad \text{sum(quantity) < 100}
\end{align*}
\]

\[Q = \text{sales(product, city, quantity)} \gamma_{\text{city, sum(quantity)}} \rightarrow \text{p} \sigma_{p < 100} \text{T2(city,p)} \rightarrow \text{T1(city,p)}\]

In class:
find a “better” plan \(P'\)
The three components of an optimizer

We need three things in an optimizer:

• Algebraic laws
• An optimization algorithm
• A cost estimator

Algebraic Laws

• Commutative and Associative Laws
  \[ R \cup S = S \cup R, \quad R \cup (S \cup T) = (R \cup S) \cup T \]
  \[ R \times S = S \times R, \quad R \times (S \times T) = (R \times S) \times T \]
  \[ R \times S = S \times R, \quad R \times (S \times T) = (R \times S) \times T \]

• Distributive Laws
  \[ R \times (S \cup T) = (R \times S) \cup (R \times T) \]
Algebraic Laws

• Laws involving selection:
  \[ \sigma_{C \text{ AND } C'}(R) = \sigma_C(\sigma_{C'}(R)) = \sigma_C(R) \cap \sigma_{C'}(R) \]
  \[ \sigma_{C \text{ OR } C'}(R) = \sigma_C(R) \cup \sigma_{C'}(R) \]
  \[ \sigma_C(R \mid \times \mid S) = \sigma_C(R) \mid \times \mid S \]

• When C involves only attributes of R
  \[ \sigma_C(R - S) = \sigma_C(R) - S \]
  \[ \sigma_C(R \cup S) = \sigma_C(R) \cup \sigma_C(S) \]
  \[ \sigma_C(R \mid \times \mid S) = \sigma_C(R) \mid \times \mid S \]

• Example: \( R(A, B, C, D) \), \( S(E, F, G) \)
  \[ \sigma_{F=3}(R \mid \times \mid D=E S) = ? \]
  \[ \sigma_{A=5 \text{ AND } G=9}(R \mid \times \mid D=E S) = ? \]
Algebraic Laws

• Laws involving projections
  \[ \Pi_M(R \times S) = \Pi_M(\Pi_P(R) \times \Pi_Q(S)) \]
  \[ \Pi_M(\Pi_N(R)) = \Pi_{M,N}(R) \]

• Example R(A,B,C,D), S(E, F, G)
  \[ \Pi_{A,B,G}(R \mid D=E S) = \Pi_{\gamma}(\Pi_{\delta}(R) \mid D=E \Pi_{\gamma}(S)) \]

Algebraic Laws

• Laws involving grouping and aggregation:
  \[ \delta(\gamma_{A, \text{agg}(B)}(R)) = \gamma_{A, \text{agg}(B)}(R) \]
  \[ \gamma_{A, \text{agg}(B)}(\delta(R)) = \gamma_{A, \text{agg}(B)}(R) \text{ if agg is “duplicate insensitive”} \]

• Which of the following are “duplicate insensitive”?
  sum, count, avg, min, max

  \[ \gamma_{A, \text{agg}(D)}(R(A,B) \mid B=C S(C,D)) = \]
  \[ \gamma_{A, \text{agg}(D)}(R(A,B) \mid B=C \gamma_{C, \text{agg}(D)}S(C,D))) \]