IEOR 115 Industrial and Commercial Data Systems
Spring 2013 Prof. Ken Goldberg

Query Optimization

An Example

Consider some relations and metadata below:

Relations

- Student(SID, SLName, Dept)
- Course(CID, Dept, CName)
- Grade(SID, CID, Sem, Grade)

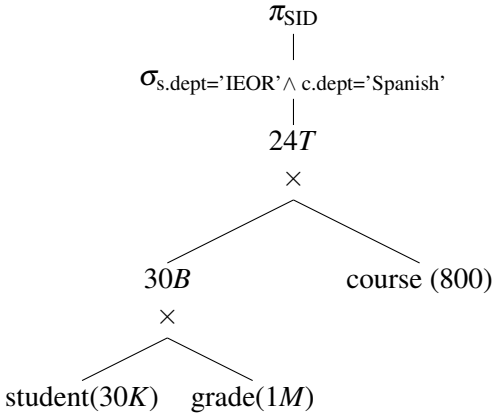
Metadata

- 30,000 students (~300 IEOR majors, ~1%)
- 800 courses (16 Spanish courses, ~2%)
- 1,000,000 grade records (~30 per student)

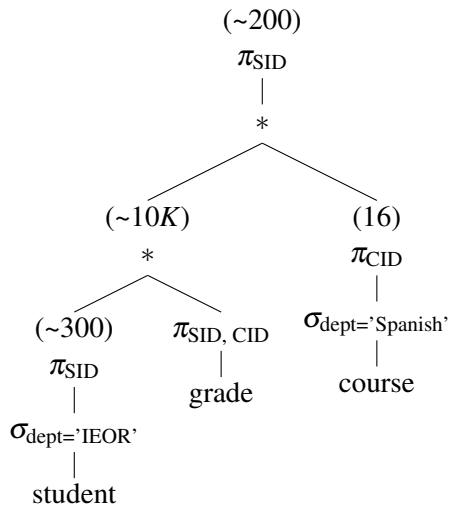
A typical SQL query would look like:

```
select distinct s.SID
from student s, course c, grade g
where s.dept = 'IEOR' and
      c.dept = 'Spanish' and
      c.CID = g.CID and
      s.SID = g.SID;
```

The query tree is shown below



Consider this optimized query tree instead:



Equivalence Rules

1. $\sigma_{c_1}(\sigma_{c_2}(...(\sigma_{c_n}(R)))) \Leftarrow \sigma_{c_1 \wedge c_2 \wedge ... \wedge c_n}(R)$ — σ - CNF Rule
   (CNF $\equiv$ Conjunctive Normal Form)
2. $\sigma_{c_2}(\sigma_{c_1}(R)) \Leftarrow \sigma_{c_1}(\sigma_{c_2}(R))$ — σ - commute Rule
3. $\pi_{Amin}(R) \Leftarrow \pi_{Amin}(\pi_{A'}(\pi_{MaxA}(R)))$ — π - minA Rule
   $Amin \subset A' \subset MaxA$
4. $R * (S * T) \Leftarrow (R * S) * T$ — $*$ - Associative Rule
5. $(T * R) * S \Leftarrow (R * S) * T$ — $*$ - Commutative Rule
6. $R - \sigma_c(R) \Leftarrow \sigma_{\neg c}(R)$ — $\neg$ Rule

CNF

- A **disjunct** is a wff that contains only $\vee$ and $\neg$ connectives. E.g. $\neg p \vee q$
  Note that formulas can be convicted into disjuncts. E.g. $p \rightarrow q \models \neg p \vee q$
- A **conjunct** is a wff that contains only $\wedge$ and $\neg$ connectives
- A **wff is in CNF** if it is a single disjunct or a set of disjuncts joined by $\wedge$.
  For example, $(\varphi_1 \vee \neg\varphi_2) \wedge (\varphi_1 \vee \varphi_3) = \{\varphi_1 \vee \neg\varphi_2, \varphi_1 \vee \varphi_3\}$

Recap: DeMorgan's Rules

- $\neg(c_1 \wedge c_2) \equiv \neg c_1 \vee \neg c_2$
- $\neg(c_1 \vee c_2) \equiv \neg c_1 \wedge \neg c_2$

Optimization Algorithm

1. For each σ put into CNF, reorder them so that the most selective conditions are applied first.
2. If you have $R \bowtie_{C_R \wedge C_S} S$, change it to $\sigma_{C_R}(R) * \sigma_{C_S}(S)$.
   Move highly selective σ as far down the query tree as possible.
3. Push the projections (π) as far down as possible.