Grundlagen Datenbanken: Übung 12

Tanmay Deshpande

Gruppe 20 & 21

ge94vem@mytum.de

QR-Code für die Folien

Wiederholung

Woche 12

Kostenbasierte Optimierung

- Wir versuchen die Auswertungsreihenfolge zu finden, die die Kosten minimiert
- Kosten werden durch Kostenfunktionen definiert (oft von der Kardinalitäten abhängig)
- Alle mögliche Kombinationen auszuprobieren ist aufwendig
- Wir brauchen Abschätzungen, die wir nutzen können
- Selektivität ist dabei hilfreich

Selektivität

- Anteil der qualifizierenden Tupel einer Operation
- Selektion mit Bedingung p:

$$sel_p := \frac{|\sigma_p(R)|}{|R|}$$

- Join von R mit S:

$$sel_{RS} := \frac{|R \bowtie S|}{|R \times S|} = \frac{|R \bowtie S|}{|R| \cdot |S|}$$

Selektivität

Abschätzung der Selektivität:

- $sel_{R.A=C} = \frac{1}{|R|}$
  falls A Schlüssel von R
- $sel_{R.A=C} = \frac{1}{i}$
  falls i die Anzahl der Attributwerte von $R.A$ ist (Gleichverteilung)
- $sel_{R.A=S.B} = \frac{1}{|R|}$
  bei Equijoin von R mit S über Fremdschlüssel in S

Ansonsten z.B. Stichprobenverfahren

Dynamisches Programmieren

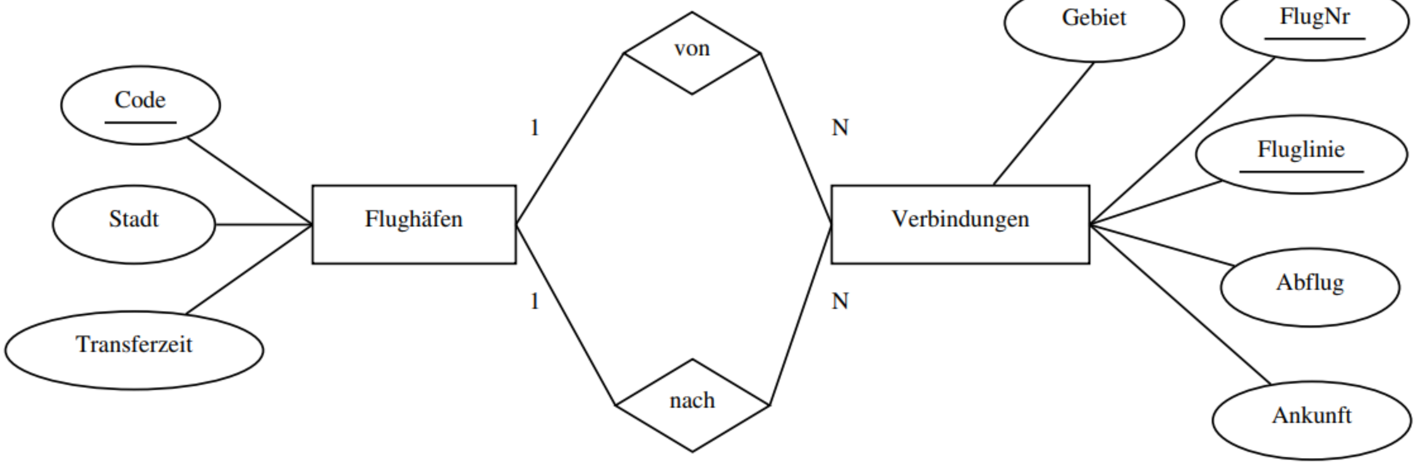
- „bottom-up“ Algorithmus
- Teile ein Problem in Teilproblemen, und versuche zuerst die kleinsten davon zu lösen
- Aufbauend auf die Lösungen der kleineren Probleme, löse Schritt-für-Schritt die größeren bis das ursprüngliche Problem gelöst ist

- Erzeuge leere DP Tabelle
  Trage Basisrelationen als optimale Lösungen der Größe 1
  ∀ Probleme der Größe s ∈ [2, n]:
      min_cost = 0
      ∀ gelösten Problemen (l, r):
          $\mathrm{l} \bowtie r$ nicht möglich oder Problemgröße(l) + Problemgröße(r) $\neq s$:
              continue
          sonst:
              cost = Kosten für $\mathrm{l} \bowtie r$
              Falls cost < min_cost: min_cost = cost
      Trage min_cost und das entsprechende Paar ein als optimale Lösung der Größe s

Aufgaben

Woche 12

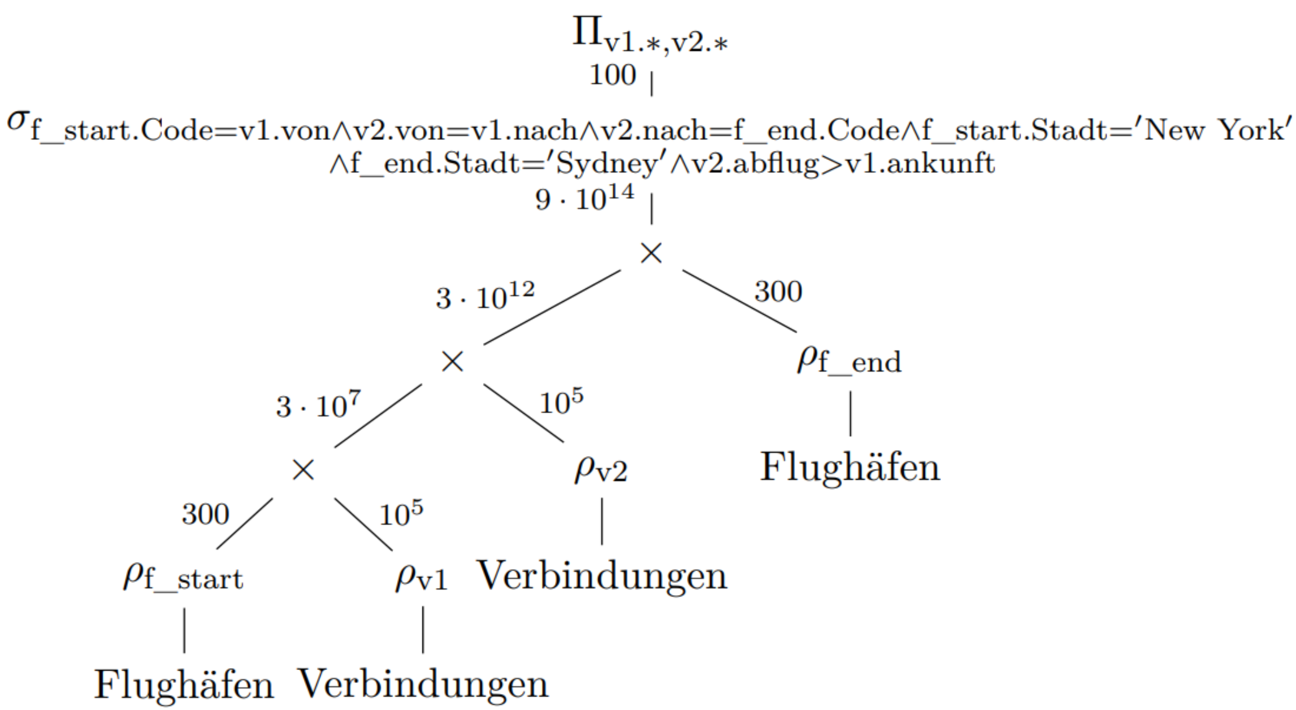
Aufgabe 01



- Finde alle Flüge von New-York nach Sydney mit einmaligem Umsteigen
- A) SQL Query für die obige Anfrage
- B) Kanonische Übersetzung der Anfrage aus A)
- C) Schätzen Sie die Relationsgrößen sinnvoll ab (z.B. so wie in den Beispielen der Vorlesung) und transformieren Sie den kanonischen Operatorbaum aus Teilaufgabe b) zur optimalen Form. Wie haben sich die Kosten dabei geändert? (Kosten = Anzahl der Zwischenergebnistupel)

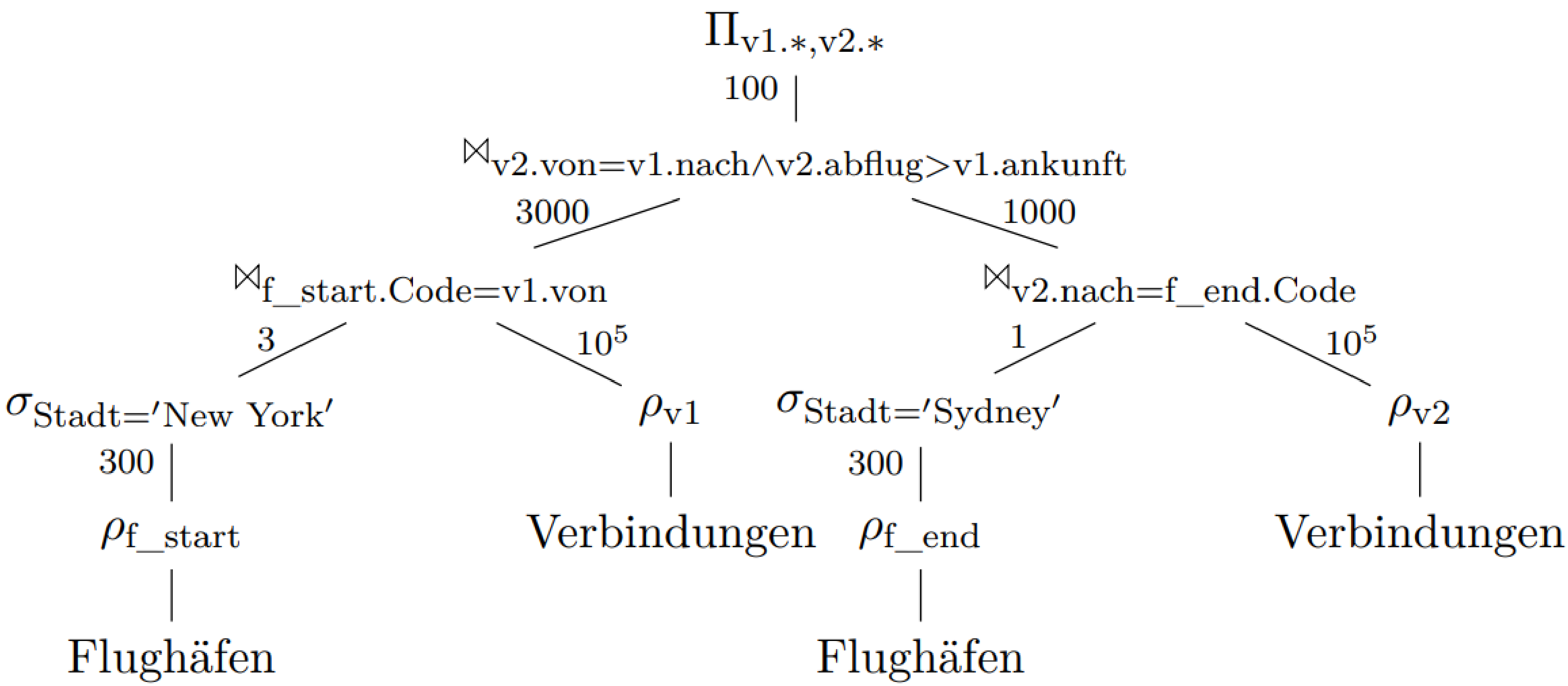
Lösungsvorschlag 01a-b

- A)
  SELECT DISTINCT v1.*, v2.* FROM
  Flughäfen f_start, Verbindungen v1, Verbindungen v2, Flughäfen f_end
  WHERE f_start.Stadt = "New York" AND f_end.Stadt = "Sydney"
  AND v2.von = v1.nach AND v2.nach = f_end.Code AND f_start.Code = v1.von
  AND v2.abflug > v1.ankunft
- B)



Lösungsvorschlag 01c

- C)



Aufgabe 02

- Für einen Join-Baum T sei folgende Kostenfunktion gegeben

$$C_{out}(T) = \begin{cases} 0 & \text{falls } T \text{ eine Basisrelation } R_i \text{ ist} \\ |T| + C_{out}(T_1) + C_{out}(T_2) & \text{falls } T = T_1 \bowtie T_2 \end{cases}$$

Die Kardinalität sei dabei

$$|T| = \begin{cases} |R_i| & \text{falls } T \text{ eine Basisrelation } R_i \text{ ist} \\ (\prod_{R_i \in T_1, R_j \in T_2} f_{i,j})|T_1||T_2| & \text{falls } T = T_1 \bowtie T_2 \end{cases}$$

- Gegeben sei eine Anfrage über die Relationen R_1, R_2, R_3 und R_4
- $|R_1|=10$, $|R_2|=20$, $|R_3| = 20$, $|R_4| = 10$
- Die Selektivitäten der Joins seien $f_{1,2} = 0.01$, $f_{2,3} = 0.5$, $f_{3,4} = 0.01$, alle nicht gegebenen Selektivitäten sind offensichtlich 1
- Vereinfachung: Joins ohne Prädikate und Operationen mit Kreuzprodukte werden nicht betrachtet
- Berechnen Sie den optimalen (niedrigste Kosten) Join-Tree

Lösungsvorschlag 2

- Wir probieren alle möglichen Kombinationen aus
- Left-Deep:

$$((R_1 \bowtie R_2) \bowtie R_3) \bowtie R_4 \tag{1}$$
$$((R_4 \bowtie R_3) \bowtie R_2) \bowtie R_1 \tag{2}$$
$$((R_3 \bowtie R_2) \bowtie R_1) \bowtie R_4 \tag{3}$$
$$((R_3 \bowtie R_2) \bowtie R_4) \bowtie R_1 \tag{4}$$

Bushy:

$$(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4) \tag{5}$$

-

$$\begin{aligned}
& C_{out}((R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)) \\
= \; & |(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)| + C_{out}(R_1 \bowtie R_2) + C_{out}(R_3 \bowtie R_4) \\
= \; & |(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)| + |R_1 \bowtie R_2| + C_{out}(R_1) + C_{out}(R_2) + |R_3 \bowtie R_4| + C_{out}(R_3) + C_{out}(R_4) \\
= \; & |(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)| + |R_1 \bowtie R_2| + |R_3 \bowtie R_4| \\
= \; & f_{1,3} \cdot f_{1,4} \cdot f_{2,3} \cdot f_{2,4} \cdot |R_1 \bowtie R_2| \cdot |R_3 \bowtie R_4| + |R_1 \bowtie R_2| + |R_3 \bowtie R_4| \\
= \; & 0.5 \cdot (0.01 \cdot 10 \cdot 20) \cdot (0.01 \cdot 20 \cdot 10) + (0.01 \cdot 10 \cdot 20) + (0.01 \cdot 20 \cdot 10) \\
= \; & 2 + 2 + 2 \\
= \; & 6
\end{aligned}$$

-

$((R_1 \bowtie R_2) \bowtie R_3) \bowtie R_4$	24
$((R_3 \bowtie R_2) \bowtie R_1) \bowtie R_4$	222
$(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)$	6

- Der optimale Baum ist der Bushy-Tree 5

Aufgabe 03

- select * from R, S, T
  where R.A = S.A and S.B = T.B and T.C = R.A

- S.A und T.C seien Fremdschlüssel auf R
- S.B sei Fremdschlüssel auf T
- R.A, T.B seien Primärschlüssel von R respektive T
- Ihre Query-Engine unterstützt nur Nested-Loop-Joins
- Kardinalitäten: |R| = 100, |S| = 1000, |T| = 10
- Es gibt keine Indexe

- a) Für Equijoins zwischen R_i mit Fremdschlüssel auf den Primärschlüssel in R_j gilt die Abschätzung:
  $$f_{i,j} = \frac{1}{|R_j|}$$
  Warum?

Aufgabe 03

- select * from R, S, T
  where R.A = S.A and S.B = T.B and T.C = R.A

- S.A und T.C seien Fremdschlüssel auf R
- S.B sei Fremdschlüssel auf T
- R.A, T.B seien Primärschlüssel von R respektive T
- Ihre Query-Engine unterstützt nur Nested-Loop-Joins
- Kardinalitäten: |R| = 100, |S| = 1000, |T| = 10
- Es gibt keine Indexe

- a) Für Equijoins zwischen R_i mit Fremdschlüssel auf den Primärschlüssel in R_j gilt die Abschätzung:

$$f_{i,j} = \frac{1}{|R_j|}$$

  Warum?

Lösungsvorschlag 03a

- Da das Datenbanksystem die referenzielle Integrität sicherstellt, referenziert jeder Fremdschlüssel mindestens ein existierendes Tupel
- Jedes Tupel aus R_i hat also mindestens einen Join-Partner in R_j . Es gibt also mindestens so viele Joinpaare wie R_i Tupel hat.
- Gleichzeitig ist das Zielattribut in R_j Primärschlüssel, also ist jeder Wert einzigartig. Für jeden Wert in R_i kann es also höchstens einen Joinpartner in R_j geben.
- Damit gibt es genau $| R_i |$ viele Joinpaare.
  Die Selektivität ist damit genau $| R_j |$, da $| R_i | \cdot | R_j | / | R_i | = 1 / | R_j |$.

Aufgabe 03

- select * from R, S, T
  where R.A = S.A and S.B = T.B and T.C = R.A

- S.A und T.C seien Fremdschlüssel auf R
- S.B sei Fremdschlüssel auf T
- R.A, T.B seien Primärschlüssel von R respektive T
- Ihre Query-Engine unterstützt nur Nested-Loop-Joins
- Kardinalitäten: |R| = 100, |S| = 1000, |T| = 10
- Es gibt keine Indexe

- b) Bestimmen Sie, wie in der Vorlesung gezeigt, den optimalen Ausführungsplan als Baum mit Kosten-/Kardinalitätsabschätzungen mit Hilfe von Dynamischem Programmieren. Verwenden Sie die Kostenfunktion C_{out}.

Lösungsvorschlag 03b

$$|R \bowtie S| = |S \bowtie R| = \frac{1}{|R|} \cdot |R| \cdot |S| = |S| = 1000$$

$$|R \bowtie T| = |T \bowtie R| = \frac{1}{|R|} \cdot |R| \cdot |T| = |T| = 10$$

$$|S \bowtie T| = |T \bowtie S| = \frac{1}{|T|} \cdot |S| \cdot |T| = |S| = 1000$$

$$|R \bowtie S \bowtie T| = \frac{1}{|R|} \cdot \frac{1}{|R|} \cdot \frac{1}{|T|} \cdot |R| \cdot |S| \cdot |T| = \frac{|S|}{|R|} = 10$$

Für die Berechnung der Kosten ergibt sich dann folgende DP-Tabelle:

DP-Tabelle		
Index	Pläne	Kosten
R	R	0
S	S	0
T	T	0
R,S	$\bowtie$ $C_{out} = 1000$ 100 / \ 1000 R S	1000
R,T	$\bowtie$ $C_{out} = 10$ 100 / \ 10 R T	10
S,T	$\bowtie$ $C_{out} = 1000$ 1000 / \ 10 S T	1000
R,S,T	$\bowtie$ $C_{out} = 1010$ 1000 / \ 100 $\bowtie$ R 1000 / \ 10 S T $\bowtie$ $C_{out} = 20$ 10 / \ 1000 $\bowtie$ S 100 / \ 10 R T $\bowtie$ $C_{out} = 1010$ 1000 / \ 10 $\bowtie$ T 100 / \ 1000 R S	20

Aufgabe 04

- select v.VorlNr, v.Titel, p.Name, count(h.MatrNr) as hoerer from
  Vorlesungen v left outer join hoeren h on (v.VorlNr = h.VorlNr), Professoren p
  where
  v.gelesenVon = p.PersNr
  group by v.VorlNr, v.Titel, p.Name
  having count(h.MatrNr) > 3

- Skizzieren Sie die kanonische Übersetzung der obigen Anfrage

Lösungsvorschlag 04