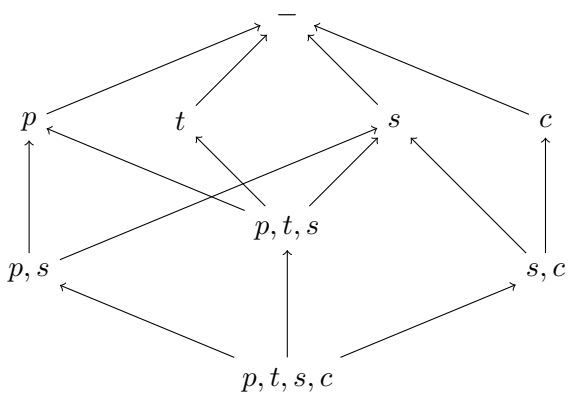
4. Übungsblatt

Besprechung: ab 29.05.

Aufgabe 1

Ein Data Warehouse-Schema enthält die vier Dimensionen p, t, s und c. Mögliche Gruppierungen und mögliche Ableitungen davon ergeben sich durch folgenden Verband (*lattice*). Die Tabelle gibt die Ergebnisgröße der möglichen Gruppierungen an.



Gruppierung	# rows
p	10
t	20
s	30
c	400
p, s	100
p, t, s	4 000
s, c	3 000
p, t, s, c	10 000

Zur Beschleunigung von Anfragen sollen materialisierte Sichten angelegt werden. Dazu ist ein (Speicherplatz-)Budget von **15 000 Tupeln** vorgesehen. Ermitteln Sie die optimale Menge von anzulegenden materialisierten Sichten. Verwenden Sie dazu das Maximum Benefit-Verfahren von Harinarayan [1], das auch in den Übungsprojekten vorgestellt wird. Notieren Sie dabei die einzelnen Schritte und protokollieren Kosten bzw. Benefits für jeden Schritt.

Aufgabe 2

Die folgende Sternschema-Anfrage soll mit einer indexbasierten Auswertungsstrategie aus der Vorlesung ausgewertet werden:

```
select sum(Sales.Revenue)
  from Sales, Territory, Date
 where Sales.TerritoryKey = Territory.Key
   and Sales.DateKey = Date.Key
   and Territory.Country = 'United States'
   and Date.CalendarYear between 2005 and 2008
```

Dazu zeigt Abbildung 1 einen Anfrageplan für die Strategie *„Index on value columns of dimension tables“* aus der Vorlesung (Strategie 1, Vorlesungsfolie 117). Die Abbildung zeigt unten vier Indizes die für den Anfrageplan genutzt werden. Der Anfrageplan geht folgendermaßen vor:

Vorgehen:

Zunächst werden unabhängig voneinander die Tupel der Dimensionstabellen *Territory* und *Date* gelesen. Dazu werden die Indizes $\texttt{IX}_1$ und $\texttt{IX}_3$ genutzt um die `rids` jener Tupel zu bestimmen, die die Filterkriterien erfüllen. Danach liest eine `FETCH`-Operation die jeweiligen Tupel.

Nun wird je Dimension ein *index nested loops join* (`INLJ`) mit der Faktentabelle berechnet. Die Indizes $\texttt{IX}_2$ und $\texttt{IX}_4$ liefern dabei jeweils die `rids` der passenden Faktentupel. Anschließend wird die Schnittmenge der beiden `rid`-Listen berechnet um die Faktentupel zu bestimmen, die sich nach beiden Dimensionen qualifizieren. Die Faktentupel werden mit einer `FETCH`-Operation gelesen und als Ergebnis aggregiert.

a) Auswertungsstrategie 1

Auf Vorlesungsfolien 120 und 121 ist ein Trick angegeben, der die Effizienz von Strategie 1 steigert. Geben Sie eine abgewandelte Form des Anfrageplans an der die Variante umsetzt. Spezifizieren Sie die verwendeten Indizes.

b) Auswertungsstrategie 2

Geben Sie einen Plan an, der die Anfrage mit Strategie 2 *„Index on primary key of dimension tables“* (Folie 118) umsetzt. Welche Indizes werden für den Plan benötigt?

c) Hub Join

Welche Indizes werden benötigt um die Anfrage mit einem Hub-Join zu verarbeiten?

d) Vergleich

Nennen Sie Vor- und Nachteile der jeweiligen Techniken.

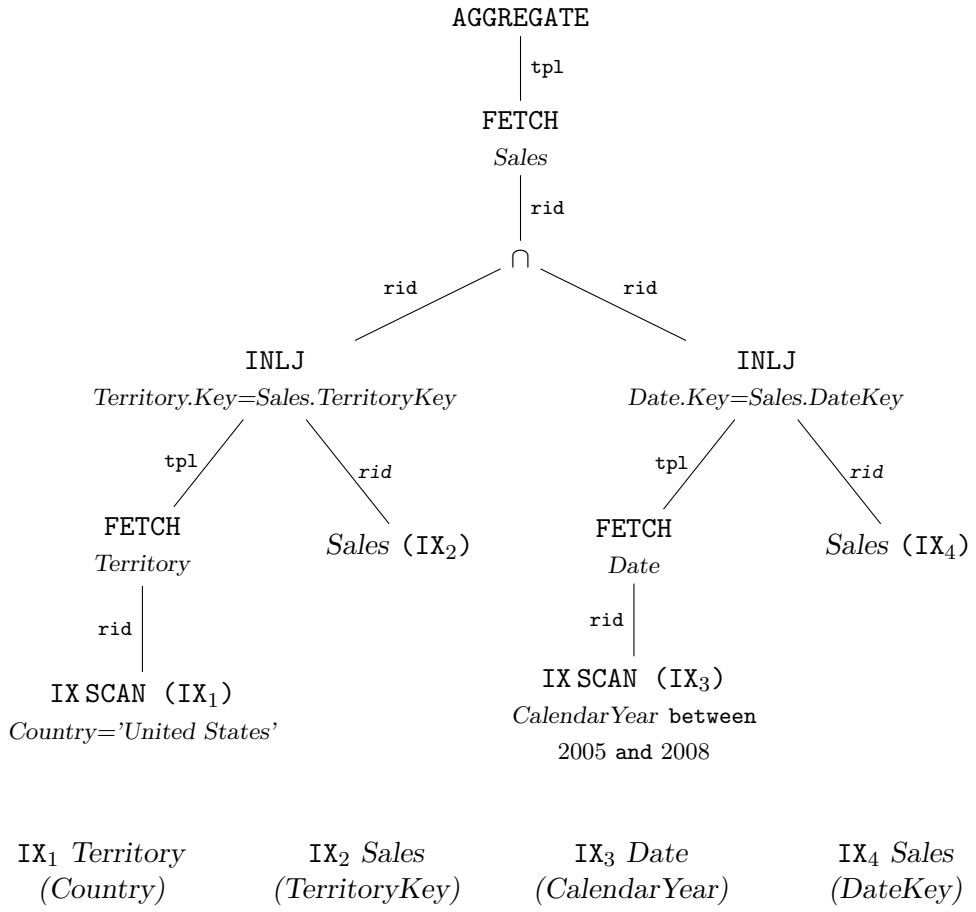


Abbildung 1: Anfrageplan für Index-Strategie 1 (Folie 117)

Literatur

[1] Venky Harinarayan, Anand Rajaraman und Jeffrey D. Ullman: Implementing Data Cubes Efficiently. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. ACM Press, 1996, Seiten 205-216.