

7. Übungsblatt

Besprechung: ab 11.07.

Aufgabe 1: Join-Verarbeitung

Betrachten Sie die folgende Star Query an ein Data Warehouse:

```
select sum(f.SalesAmount) as Betrag ,  
        dt.SalesTerritoryRegion as Verkaufsgebiet ,  
        dp.ModelName as Model ,  
        dd.CalendarYear as Jahr  
from FactResellerSales f  
inner join DimSalesTerritory dt  
        on f.SalesTerritoryKey=dt.SalesTerritoryKey  
inner join DimProduct dp on f.ProductKey = dp.ProductKey  
inner join DimReseller dr on f.ResellerKey=dr.ResellerKey  
inner join DimDate dd on f.OrderDateKey=dd.DateKey  
where dt.SalesTerritoryCountry = 'United_States'  
        and dd.CalendarYear between 2005 and 2008  
        and dr.NumberEmployees>=100  
group by SalesTerritoryRegion , dd.CalendarYear , dp.ModelName
```

In der Vorlesung haben Sie zwei unterschiedliche Strategien zur Anfrageausführung kennengelernt. Erklären Sie die Strategien an der gegebenen Anfrage! Erläutern Sie welche Indexstrukturen bei den unterschiedlichen Strategien angewendet werden! Nennen Sie die Vor- und Nachteile der jeweiligen Strategien!

Aufgabe 2: Map Reduce

MapReduce ist ein Programmiermodell, bei dem der Programmierer einen Mapper als Funktion $f_1 : \alpha \rightarrow [(\beta, \gamma)]$ definiert, die aus einer Eingabe α eine Liste von Schlüssel-Wert Paaren $\langle \beta, \gamma \rangle$ extrahiert, und einen zugehörigen Reducer als Funktion $f_2 : \langle \beta, [\gamma] \rangle \rightarrow \delta$, die für einen Schlüssel β eine Liste aller zugehörigen Werte γ verarbeitet. Eine Implementierung dieses Modells, z.B. Hadoop, instanziiert dann diese beiden Funktionen automatisch auf verteilten Maschinen, verteilt eine Liste $[\alpha]$ von Eingaben auf die Mapper Instanzen und überführt die Ausgaben der Mapper Instanzen in einem Shuffle Schritt in die passenden Eingaben der Reducer Instanzen.

Sie sollen nun die folgende Anfrage mittels MapReduce berechnen lassen:

```
select f.DateKey, sum(f.SalesAmount)
from FactResellerSales f
group by f.DateKey
```

Die Berechnung sollen Sie in folgenden Teilaufgaben beschreiben:

1. Definieren Sie die Funktionen f_1 (Mapper) und f_2 (Reducer) durch Pseudocode.
2. Zeigen Sie anhand einfacher Beispieldaten, wie MapReduce auf $n \geq 4$ Maschinen mit Ihren Funktionen die gegebene Anfrage auswertet. Orientieren Sie sich bei der Darstellung an dem Beispiel auf Folie 202 der Vorlesung (Distributed Index Generation).