Übungsprojekt 4

Slice Operation mit MapReduce

MapReduce ist ein Programmiermodell, bei dem der Programmierer einen Mapper als Funktion $\mathsf{f}_1 : \alpha \to [\langle \beta, \gamma \rangle]$ definiert, die aus einer Eingabe α eine Liste von Schlüssel-Wert Paaren $\langle \beta, \gamma \rangle$ extrahiert, und einen zugehörigen Reducer als Funktion $\mathsf{f}_2 : \langle \beta, [\gamma] \rangle \to \delta$, 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.

In diesem Übungsprojekt sollen Sie eine Slice Operation mittels MapReduce berechnen lassen. Das Übungsprojekt besteht dabei aus den folgenden Teilaufgaben:

1. Definieren Sie die Funktionen f_1 (Mapper) und f_2 (Reducer) durch Pseudocode zunächst für eine einfache Slice Operation in der Form $\mathit{Fakt} \bowtie \sigma(\mathit{Dim1})$. Überlegen Sie, ob die f_2 Funktion außer dem Schlüssel β noch Hilfsdaten benötigt. Diese können Sie zusätzlich in den Wert γ schreiben. Für Ihre Lösung können Sie gerne auch andere Quellen zur Hilfe nehmen, z.B. Online Quellen.

2. Zeigen Sie anhand einfacher Beispieldaten, wie MapReduce auf $n \geq 4$ Maschinen mit Ihren Funktionen die folgende Anfrage auswertet:

   ```
   select *
   from   FactResellerSales f  inner join DimSalesTerritory dt
                on f.SalesTerritoryKey=dt.SalesTerritoryKey
   where dt.SalesTerritoryCountry = 'United States'
   ```

   Dabei soll eine Maschine eine Instanz (Mapper oder Reducer) ausführen. Orientieren Sie sich dabei bei der Darstellung an dem Beispiel auf der Vorlesung (Distributed Index Generation).

3. Skizzieren Sie, wie Ihre Lösung auf mehrfache Slice Operationen erweitert werden könnte wie für die folgende Beispiel Anfrage, und diskutieren Sie Nachteile Ihrer Erweiterung sowie mögliche Alternativen.

   ```
   select *
   from   FactResellerSales f  inner join DimSalesTerritory dt
                on f.SalesTerritoryKey=dt.SalesTerritoryKey
   ```

```
                    inner join DimDate dd on f.OrderDateKey=dd.DateKey
where dt.SalesTerritoryCountry = 'United States'
and dd.CalendarYear=2008
```