technische universität
dortmund

DATENBANKEN IN DER PRAXIS:
DATA WAREHOUSING
Wintersemester 2015/2016
Prof. Dr. Jens Teubner · DBIS Group
Übung: Dr. Cornelia Tadros · ISSI Group

# 7. Übungsblatt

Besprechung: 2.2.2016 (Gruppe A, 14:15 Uhr; Gruppe A+B 16:15 Uhr),
9.2.2016 (Gruppe B, 14:15 Uhr)

## Aufgabe 1

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 $\alpha$ 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 $\beta$ eine Liste aller zugehörigen Werte $\gamma$ 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 $\mathsf{f}_1$ (Mapper) und $\mathsf{f}_2$ (Reducer) durch Pseudocode.

2. Zeigen Sie anhand einfacher Beispieldaten, wie MapReduce auf $n \geq 4$ Maschinen mit Ihren Funktionen die gegebene Anfrage auswertet. Dabei soll eine Maschine eine Instanz (Mapper oder Reducer) ausführen. Orientieren Sie sich bei der Darstellung an dem Beispiel auf Folie 202 der Vorlesung (Distributed Index Generation).