7. Übungsblatt

Besprechung: ab 10.07.

Aufgabe 1: Gruppierte Aggregation mit MapReduce

MapReduce ist ein Programmiermodell für Clusteranwendungen. Der Programmierer spezifiziert eine Anwendung durch eine Map-Funktion und eine Reduce-Funktion:

Map $\mathsf{f}_1 : \alpha \to [\langle \beta, \gamma \rangle]$
Map extrahiert aus einer Eingabe α eine Liste von Schlüssel-Wert Paaren $\langle \beta, \gamma \rangle$.

Reduce $\mathsf{f}_2 : \langle \beta, [\gamma] \rangle \to \delta$
Reduce verarbeitet für einen Schlüssel β die Liste aller zugehörigen Werte γ.

Systeme wie z.B. Hadoop führen den Map und Reduce Schritt dann auf einem verteilt gespeicherten Datensatz aus. Zwischen Map und Reduce sorgt *Shuffle* dafür, dass alle Elemente mit dem gleichem Schlüssel an den selben Reduce Aufruf geleitet werden.

Sie sollen nun die folgende Anfrage mittels MapReduce berechnen lassen:

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

Die Berechnung sollen Sie in folgenden Teilaufgaben beschreiben:

1. Definieren Sie die Funktionen f_1 (Map) und f_2 (Reduce) durch Pseudocode.
2. Zeigen Sie anhand einfacher Beispieldaten, wie MapReduce auf $n \geq 4$ Maschinen mit Ihren Funktionen die gegebene Anfrage auswertet.

Aufgabe 2: MapReduce Verarbeitungsmodelle

Für Clusteranwendungen haben sich unterschiedliche Verarbeitungsmodelle etabliert. Zum Beispiel wird das klassische MapReduce Modell [1] durch Apache Hadoop umgesetzt. Apache Spark [2] erweitert MapReduce und verspricht einige Verbesserungen.

Erklären auf Basis der Quellen welche Vorteile Spark gegenüber Hadoop bietet. Gehen Sie dabei auf die folgenden Aspekte ein:

a) Datenfluss

Welche Erweiterungen des Datenflusses bietet Spark gegenüber dem Ablauf Map → Reduce?

b) Datenhaltung

Wie ermöglicht Spark schnellere Datenzugriffe bei mehrstufigen Datenflüssen?

Literatur

[1] Jeffrey Dean et al.: MapReduce: simplified data processing on large clusters. Communications of the ACM 2008.

[2] Matei Zaharia et al.: Spark: Cluster computing with working sets. HotCloud 2010.