Aufgabenblatt 7
Web-Scale Data Management

- Abgabetermin: **-** (dieses Blatt wird weder abgegeben noch bewertet)
- Dieses Aufgabenblatt enthält einige Übungen zum Thema Web Scale Data Management.

Aufgabe 1: MapReduce

Ein optisches Messverfahren erfasst die Oberfläche von Objekten als Menge von 3D-Ortsvektoren $(x, y, z)^T$ und speichert diese in der Datenbanktabelle $S(\underline{vectorId}, \underline{dimension}, value)$ ab. Durch Störeinflüsse werden beim Messen nicht alle Ortsvektoren vollständig erfasst.

Für eine grafische Simulation werden nun die Vektorlängen *length* aller **vollständig erfasster** Ortsvektoren benötigt, $length = \sqrt{x^2 + y^2 + z^2}$. Aufgrund der großen Datenmenge soll die Berechnung auf einem MapReduce-Cluster erfolgen.

a) Als Ausgabe soll eine Tabelle $T(\underline{vectorId}, length)$ erzeugt werden, in der für jeden vollständig erfassten Ortsvektor die Vektorlänge steht. Die Ausgabe wird partitioniert auf den verschiedenen Nodes ins verteilte Dateisystem geschrieben. Lösen Sie die Aufgabe mit nur einem MapReduce-Job. Verwenden Sie auch die in der Übung vorgestellte Funktion `combine`, um Netzwerklast zu reduzieren. Beschreiben Sie Ihr Vorgehen kurz in wenigen Sätzen. Erstellen Sie den Pseudocode für `map`, `combine` und `reduce`. **8 P**

b) Die Abbildung zeigt beispielhaft eine Verteilung von S auf drei MapReduce-Nodes. Zeigen Sie für dieses Beispiel die Ausgabe der einzelnen Phasen (`map`, `combine` und `reduce`) Ihres MapReduce-Jobs auf jeder der drei Nodes. **3 P**

id	*dim*	*val*
1	x	2
2	z	4
1	y	3
3	y	2
4	y	4

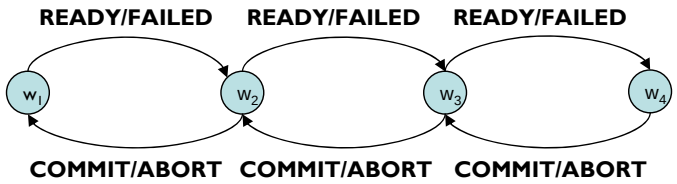
id	*dim*	*val*
3	z	4
4	z	2
2	x	2
2	y	4

id	*dim*	*val*
5	y	3
3	x	4
4	x	4

Aufgabe 2: Two Phase Commit (Bonusaufgabe)

In der Vorlesung wurde eine hierarchische Organisationsstruktur (ein Koordinator und mehrere untergeordnete Worker) beim 2PC-Prototoll beschrieben. Es ist auch möglich, die in der Abbildung gezeigte lineare Organisationsstruktur vorzunehmen.

Hierbei ist kein ausgezeichneter Koordinator erforderlich. In der ersten Phase reichen die Worker ihren eigenen Status und den der linken Nachbarn von *links nach rechts* weiter, nachdem sie einen entsprechenden Statusbericht von links bekommen haben. Der letzte Worker in der Reihe – hier Worker w_4 – trifft die Entscheidung und reicht sie nach links weiter.



Entwickeln Sie das 2PC-Protokoll – analog zur Vorlesung – für diese lineare Anordnung der Worker als Pseudocode und beschreiben Sie kurz das Vorgehen. **4**
Bonuspunkte