Praktikum 4: Datenanalyse mit Hadoop / MapReduce

Ziel des Praktikums

Ziel dieses Praktikums ist es, sich mit Hadoop vertraut zu machen. Im nächsten Praktikum wird der Schwerpunkt auf der Skalierung mit Hadoop liegen.

Aufgabe 1 (Arbeit mit Hadoop – Vorbereitung zu Hause)

Machen Sie sich mit Hadoop vertraut. Eine Anleitung zur Verwendung der Hadoop-Installation auf dem Cluster und Beispiele zur grundsätzlichen Arbeit mit Hadoop finden Sie im Wiki. Testen Sie das WordCount-Beispiel (siehe Wiki).[1]

Aufgabe 2 (Vorbereitung der Daten)

Die untenstehenden Anfragen sollen jeweils auf 1, 10 bzw. 20 Millionen Datensätzen ausgeführt werden. Verwenden Sie die aufbereiteten und schon bekannten Daten unter `.../JSON`. Legen Sie die Daten geeignet im HDFS ab, damit sie mit Hadoop verarbeitet werden können.

Aufgabe 3 (Datenbankabfragen mit MapReduce)

Implementieren Sie anschließend die untenstehenden Abfragen mit MapReduce auf Hadoop und messen und dokumentieren Sie die Zeiten auf den unterschiedlichen Datenmengen. Bemühen Sie sich um eine effiziente Umsetzung und diskutieren Sie im Praktikumsbericht ggf. verschiedene Varianten!

1. Ausgabe der Anzahl aller Ratings.
2. Ausgabe aller Filmtitel, die der Nutzer mit der ID = 10 bewertet hat.
3. Ausgabe der Anzahl aller Ratings zu jedem Nutzer.
4. Ausgabe aller Filme mit einer durchschnittlichen Bewertung $>= 4$.

Hinweis: Verwenden Sie zur Umsetzung beispielsweise den `JSONParser` aus dem Package `org.json.simple`. Eine funktionierende `json-simple-1.1.1.jar` Library finden Sie bereits auf dem Cluster unter `/mnt/datasets/libs`. Sie können entweder die `jar`-Datei lokal erzeugen und die benötigten Libraries einbinden und dann das `jar` auf dem Cluster ausführen. Oder Sie verwenden das Tool Interface (Klasse `ToolRunner`)[2], mit dem dynamisch Libraries eingebunden werden können.

Alternativ stellen wir Ihnen eine Rahmenimplementierung für die erste Abfrage inklusive Hinweisen zur Verwendung im Moodle zur Verfügung.

[1] Achten Sie bei anderen Beispielen und Tutorials darauf, dass Sie Beispiele für die aktuelle auf dem Big Data Cluster installierte Variante wählen (keine Beispiele für Hadoop 1.x).

[2] `https://hadoopi.wordpress.com/2013/06/05/hadoop-implementing-the-tool-interface-for-mapreduce-driver/`

Praktikumsbericht

Abgabe des Praktikumsberichts eine Woche nach dem jeweiligen Praktikumstermin.