MapReduce

Datenverarbeitungsverfahren zur verteilten Analyse von riesigen Datenmengen.

Marvin Heimbrodt – FSC81 – AS – 28.02.19

Überblick

MapRecude ist ein von Google entwickeltes Programmiermodell. Es folgt dem einfachen Prinzip „Aufgaben vom selben Typ lassen sich schneller verarbeiten“. Der Mehraufwand, durch vorherige Sortierung der Aufgaben, ist dabei vernachlässigbar, da MapReduce für Datenmengen im Petabyte-Bereich genutzt werden kann.

Verarbeitungsphasen

Ausgangspunkt sind jeweils eine große Menge an unstrukturierten Daten, die an das MapReduce-System übergeben wird.

Map

Die vorher unstrukturierten Daten werden in Key/Value-Paare zerteilt. Welcher Teil der Daten Key und Value sind, wird bereits hier, abhängig vom benötigten Endergebnis, gewählt.

Shuffle

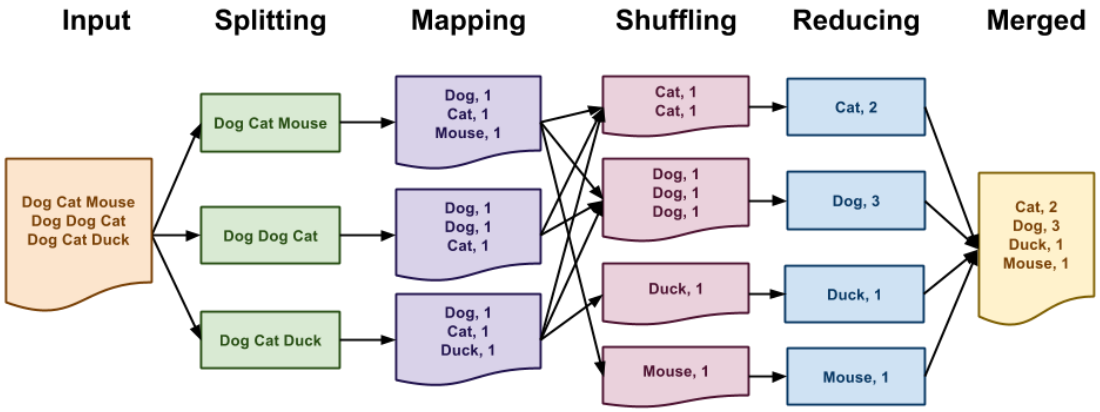
Die Key/Value-Paare werden sortiert und gruppiert. Gleiche Paare werden demselben Rechenprozess zugeordnet.

Reduce

Jeder Rechenprozess berechnet jetzt nur das Zwischenergebnis für seine Daten.

Die Zwischenergebnisse werden jetzt zusammengefasst und als Endergebnis zurückgegeben.

Beispiel mit Wortzählug



1. Peter Bui@ University of Notre Dame
https://www3.nd.edu/~pbui/teaching/cse.30331.fa16/

Apache Hadoop

Apache Hadoop ist eine konkrete Implementierung des MapReduce Verfahren. Hadoop bietet ein Framework, das die einzelnen Phasen von MapRecude auf viele Computer bzw. Nodes aufteilt. In der Shuffle-Phase werden die Key/Value-Paare dann jeweils auf eine Node bzw. einen Computer gruppiert. Der Programmierer liefert Mapper, Reducer und die Quelldaten – Hadoop koordiniert die Verarbeitung auf den Nodes und liefert ein Endergebnis.

Hadoop nutzt auf den Nodes das „Hadoop Distributed File System“ (HDFS) als verteiltes Dateisystem. Die Quelldaten werden damit automatisch auf den Nodes dupliziert – so reduziert sich das Datenvolumen in der Shuffle-Phase und der Anspruch bzgl. Daten-/Ausfallsicherheit.

Beispiel mit Java: https://www.guru99.com/create-your-first-hadoop-program.html