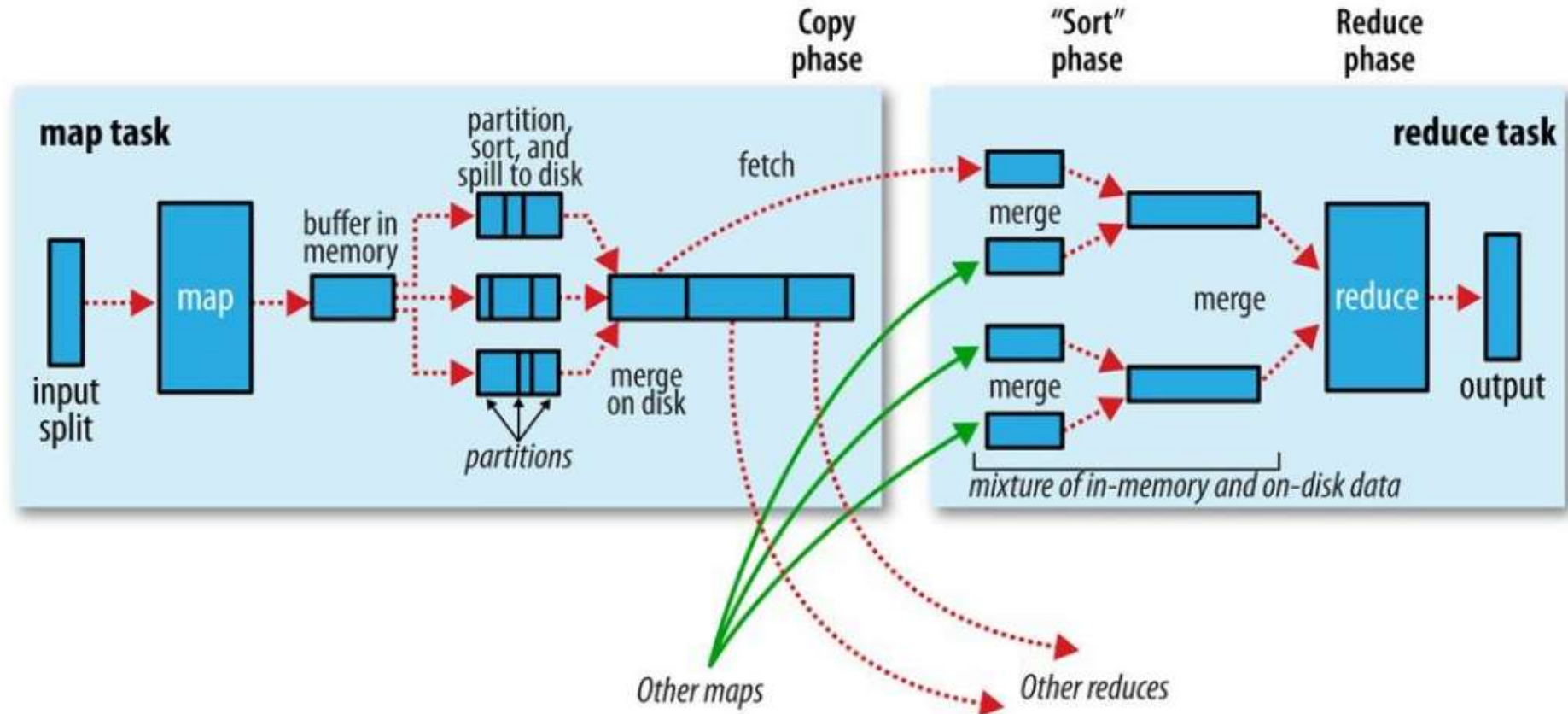


# Übersicht



Quelle: Hadoop – The Definitive Guide

# MapReduce

## 1. Map

- Mappers map input data to intermediate key/value pairs parse, filter, or transform the data
- Each Map task operates on a single HDFS block
- run on the node where the block is stored

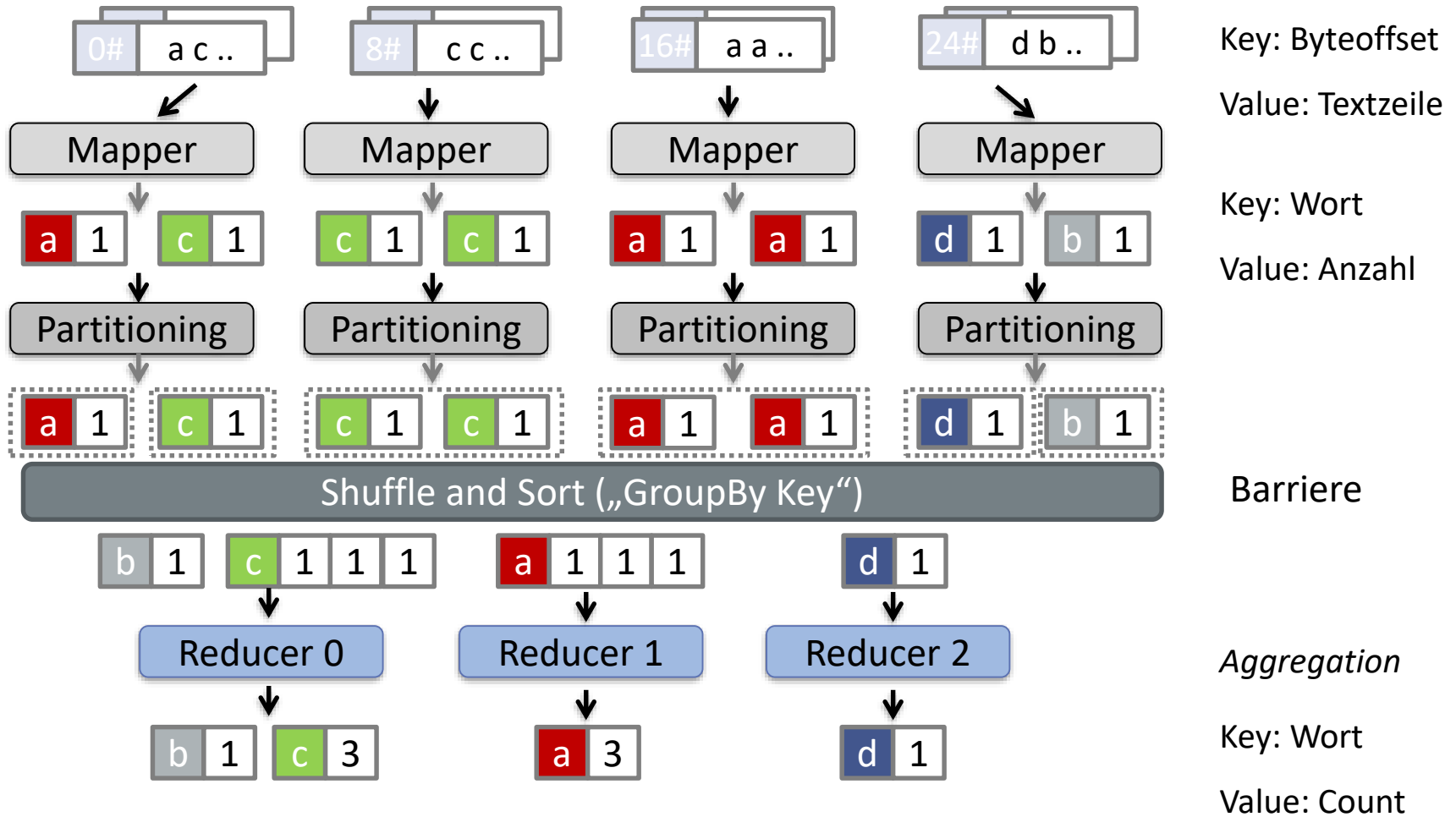
## 2. Shuffle and Sort

- Sorts and consolidates intermediate data from all mappers
- Happens after all Map tasks are complete and before Reduce tasks start → place for optimization

## 3. Reduce

- Reducers process Mapper output into final key/value pairs aggregate data using statistical functions
- Operates on Map task output (shuffled/sorted intermediate data)

# MapReduce: WordCount



# MapReduce-Optimierung: Combiner

