Proposal

Jos van der Til & Rene Zuidhof

March 21, 2011

Comparison of Map/Reduce implementations

MapReduce is a software framework introduced by Google to support distributed computing on large data sets using clusters of computers. Using MapReduce large data sets can be processed in a short amount of time in two steps. The first step is splitting the main task into multiple smaller subtasks, called mapping. These subtasks are then processed on the computers in the cluster.

The subtasks produce partial answers which are then gathered and combined to produce the final solution, this is second step called reducing. This concept can be compared to the divide & conquer concept used in other algorithms, for example MergeSort, and is therefore not entirely new. Using this on a larger scale and over a cluster of computers, a new dimension of data processing is reached with its own problems and difficulties.

Examples of MapReduce usage are the indexing of large amounts of documents or analyzing visitor behavior by Google.

The proposed paper will begin with an introduction to the basics and the needs of MapReduce implementations. Followed by discussions of the different implementations (e.g., Hadoop, CouchDB). At the end the different implementations will be analyzed in terms of performance, applicability, scalability etc. after which the results of some experiments done by ourselves will be discussed.

The main goal is to present the pros & cons of the different implementations.

Field of research: Software Engineering
Topic: Comparison of Map/Reduce implementations
Focus/research question: What is the best MapReduce implementation
Expected findings/results: Find the most important differences between multiple MapReduce implementations and what would work best in what cases.