High-performance log data analysis using MapReduce

Fernand Geertsema (s1947427)
Erwin Vast (s1924451)

Abstract - MapReduce is a framework patented by Google for processing large amounts of data. The processing of the data is divided in multiple subtasks, which are distributed over multiple servers. The MapReduce framework can be used for different problems, like counting word occurrences or distributed sorting. Our research is aimed at log analysis of server-parks. Most existing solutions process the log data of server-parks in batches, however, to monitor them real-time it is necessary to process a continues stream of log data.

In our research we develop an application that analyzes such a data stream using MapReduce. Each server logs, for example, the processor usage, memory usage and network throughput. By analyzing this log data, it is possible to extract the performance and status of the entire server-park. The analysis will consist of averaging all the server logs to one log for the complete server-park. In our paper we describe what a possible solution is to do log data analysis using MapReduce.

The performance of the application is very important, because the amount of log data for a server-park will be huge. Based on this performance requirement a MapReduce implementation will be chosen that can be used by our application. There are multiple MapReduce implementations available like Hadoop or CouchDB. A simulated server-park generates the data that is processed by our application. From this data the performance and status of the server-park are extracted.

Keywords: MapReduce, high-performance, distributed computing, log data analysis, implementations, comparison, Hadoop.
Field of research: Distributed computing (data reduction)
Topic: Log data analysis using MapReduce
Research question: What is possible solution to do high-performance log data analysis using MapReduce?