Abstract

This paper studied the design, implementation and evaluation of a MapReduce tool targeting distributed systems, and multi-core system architectures. MapReduce is a distributed programming model originally proposed by Google for the ease of development of web search applications on a large number of clusters of computers. We addressed the issues of limited resource for data optimization for efficiency, reliability, scalability and security of data in distributed, cluster systems with huge datasets. The study’s experimental results predicted that the MapReduce tool developed improved data optimization. The system exhibits undesired speedup with smaller datasets, but reasonable speedup is achieved with a larger enough datasets that complements the number of computing nodes reducing the execution time by 30% as compared to normal data mining and processing. The MapReduce tool is able to handle data growth trendily, especially with larger number of computing nodes. Scaleup gracefully grows as data and number of computing nodes increases. Security of data is guaranteed at all computing nodes since data is replicated at various nodes on the cluster system hence reliable. Our implementation of the MapReduce runs on distributed cluster computing environment of a national education web portal and is highly scalable.
Use of Mapreduce for Data Mining and Data Optimization on a Web Portal

References


Index Terms

Computer Science  Information Sciences

Keywords

MapReduce  Hadoop  Scalability