Solar Energy | Water | Environment
Nexus
IN NEVADA

Advanced MapReduce Process Using Limited Node Block Placement Policy

Sungchul Lee[1], Ju-Yeon Jo[2], Yoohwan Kim[3]

Department of Computer Science
University of Nevada Las Vegas, Las Vegas, NV 89154
[1]Email: lees173@unlv.nevada.edu
[2, 3]Email: {Juyeon.Jo, Yoohwan.Kim}@unlv.edu

MapReduce has been widely used in many data science applications. To improve its performance, we have studied the processes of Map and Shuffle, and identified inefficiency associated with Rack-Local Map (RLM). It has been also observed that an excessive data transfer during the shuffle process has a negative impact on the performance. In this research, we introduce a new block placement paradigm called Limited Node Block Placement Policy (LNBPP). Under the conventional default block placement policy (DBPP), data blocks are randomly placed on any available slave nodes, requiring a copy of RLM blocks. On the other hand, LNBPP places the blocks in a way to avoid RLMs, reducing the block copying time. The containers without RLM has a more consistent execution time, and when assigned to individual cores on a multicore node, they finish a job faster collectively than the containers under DBPP. LNBPP also rearranges the blocks into a smaller number of nodes (hence Limited Node) and minimizes the data transfer overhead between nodes. These strategies bring a significant performance improvement in Map and Shuffle processes. Our test results show that the execution time of Map and Shuffle can be improved by up to 33% and 44% respectively. In this paper, we describe the MapReduce workflow in Hadoop with a simple computational model and introduce the current research directions in each step on improving the performance of MapReduce. We analyze the block placement status and RLM locations in DBPP with the customer review data from TripAdvisor and measure the performances by executing the Terasort Benchmark with various sizes of data.

This material is based upon work supported by the National Science Foundation under Grant No. IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.