English

Research Express@NCKU Volume 30 Issue 5 - May 6, 2016
[http://research.ncku.edu.tw/re/articles/e/20160506/4.html]

A dynamic data placement strategy for Hadoop in heterogeneous environments

Chia-Wei Lee[1], Kuang-Yu Hsieh[1], Sun-Yuan Hsieh[1,2,*], and Hung-Chang Hsiao[1]

[1] Department of Computer Science and Information Engineering, National Cheng Kung University
[2] Institute of Manufacturing Information Systems, National Cheng Kung University

hsiehsy@mail.ncku.edu.tw

Big Data Research, (special issue on Scalable Computing for Big Data), vol. 1, pp. 14-22, August 2014.

Cloud computing is a type of parallel distributed computing system that has become a frequently used computer application. MapReduce is an effective programming model used in cloud computing and large-scale data-parallel applications. Figure 1 is the overview of the MapReduce model.

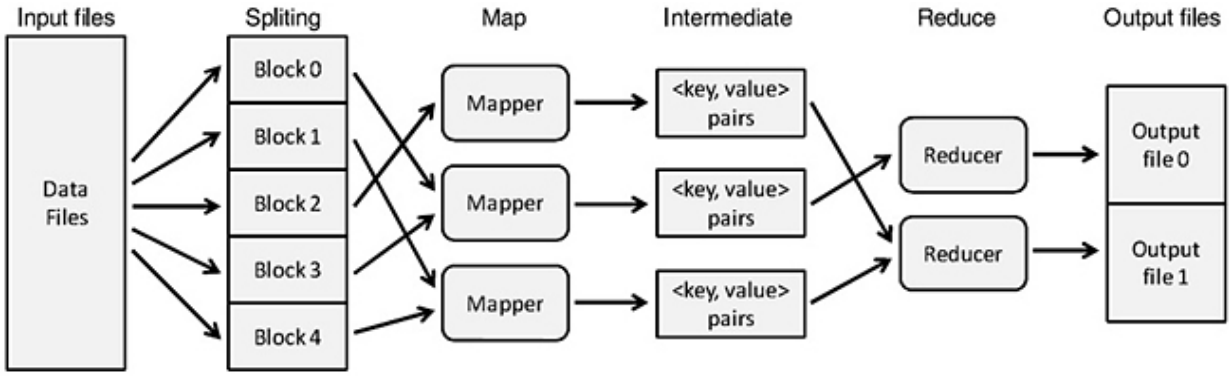

Figure 1. The overview of the MapReduce model.

Hadoop is an open-source implementation of the MapReduce model, and is usually used for data-intensive applications such as data mining and web indexing. The current Hadoop implementation assumes that every node in a cluster has the same computing capacity and that the tasks are data-local, which may increase extra overhead and reduce MapReduce performance. Figure 2 shows the default data allocation strategy of Hadoop.

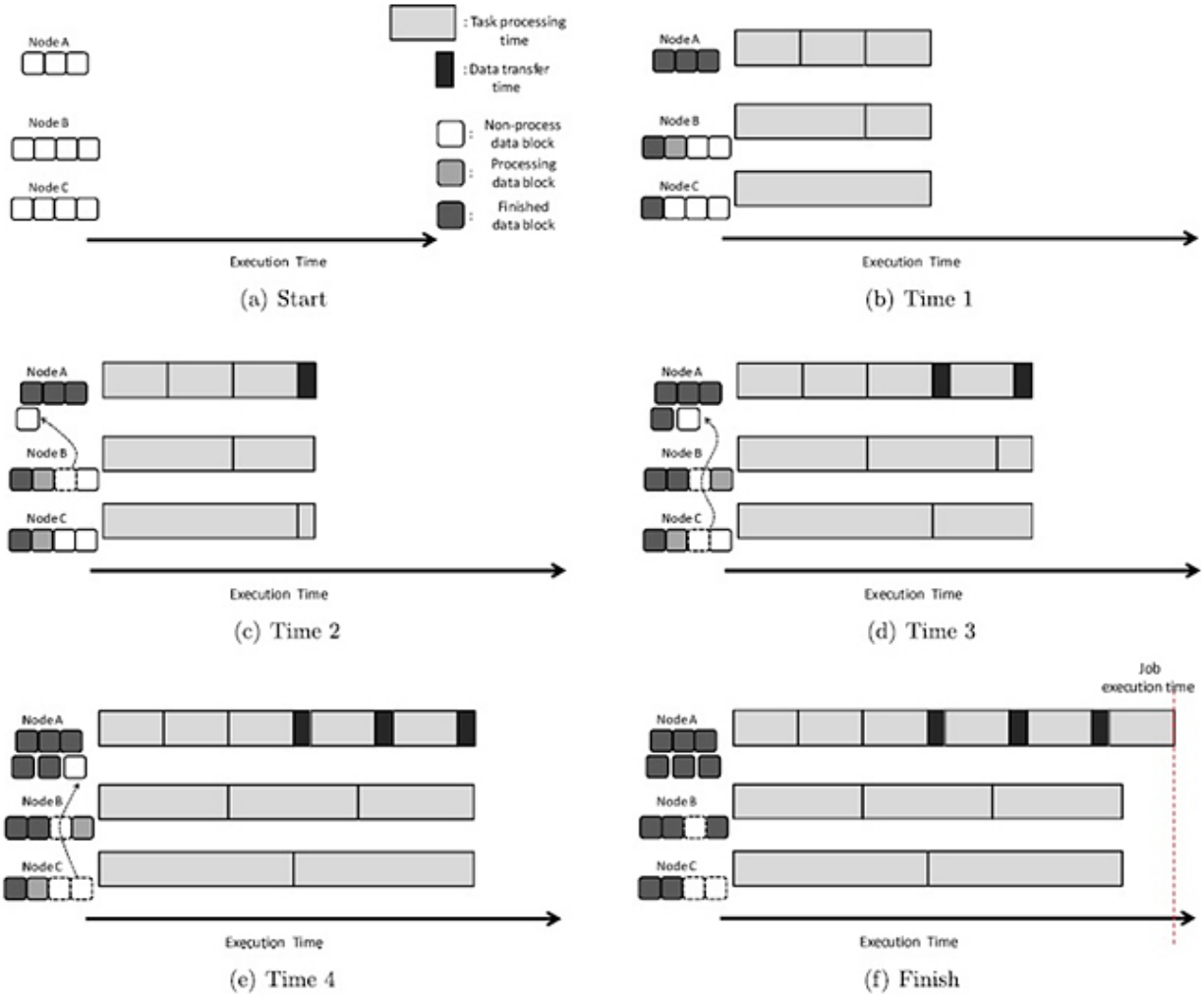

Figure 2. The default data allocation strategy of Hadoop.

This paper proposes a data placement algorithm to resolve the unbalanced node workload problem. The proposed method can dynamically adapt and balance data stored in each node based on the computing capacity of each node in a heterogeneous Hadoop cluster. The proposed method can reduce data transfer time to achieve improved Hadoop performance. Figure 3 shows the best case of data allocation.

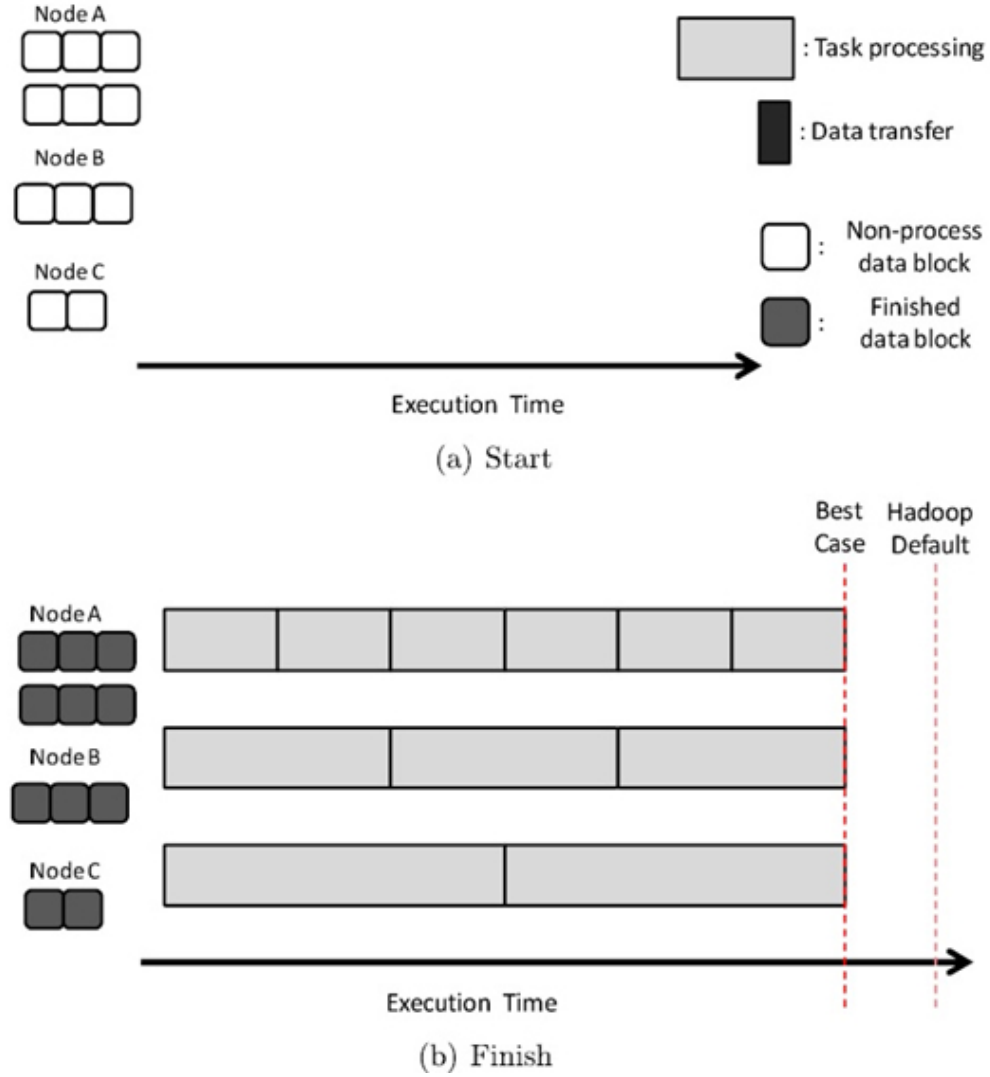

Figure 3. The best case of data allocation.

The experimental results show that the dynamic data placement policy can decrease the time of execution and improve Hadoop performance in a heterogeneous cluster. Figure 4 shows the experimental results.

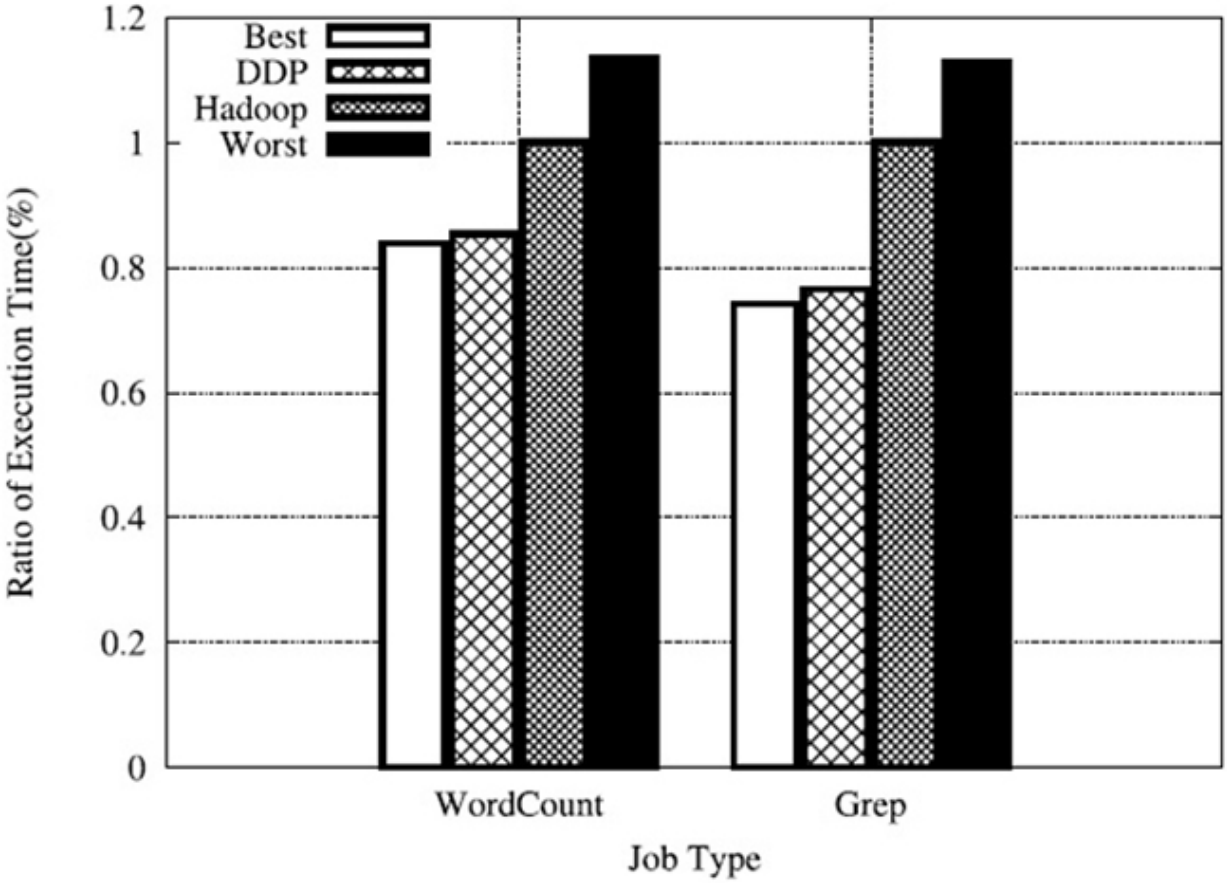

Figure 4. The percent of average execution time of a job compare with Hadoop default strategy.