

Network-Aware Multiway Join for MapReduce
Kenn Slagter, 許慶賢, Yeh-Ching Chung, Jong Hyuk Park
Computer Science & Information Engineering
Computer Science and Informatics
chh@chu.edu.tw

Abstract

MapReduce is an effective tool for processing large amounts of data in parallel using a cluster of processors or computers. One common data processing task is the join operation, which combines two or more datasets based on values common to each. In this paper, we present a network aware multi-way join for MapReduce (NAMM) that improves performance by redistributing the workload amongst reducers. NAMM achieves this by redistributing tuples directly between reducers with an intelligent network aware algorithm. We show that our presented technique has significant potential to minimize the time required to join multiple datasets.

Keyword : MapReduce, Hadoop, Multiway Join, Workload Redistribution.