Distributed Computing and Hadoop in Statistics

Xiaoling Lu and Bing Zheng
Center For Applied Statistics, Renmin University of China, Beijing, China
xiaolinglu@ruc.edu.cn；jlice1026@163.com

Big data is ubiquitous today. The big data challenges current numeric statistical and machine learning methods, visualization methods, computational methods and computational environments. Distributed Computing and Hadoop help solve these problems. Hadoop is an open source framework for writing and running distributed applications. It consists of the MapReduce distributed compute engine and the Hadoop Distributed File System (HDFS). Mahout produces machine-learning algorithms on the Hadoop platform. Rhadoop, Rhive and RHipe are the merger of R and Hadoop. Finally we show two examples of how enterprises use Hadoop to their big data processing problems.

Keywords: Distributed Computing, Hadoop, Statistics, Mahout, R