Abstract

Big data has gained its popularity in the recent years due to the fact that there is a need for sophisticated method to collect, process, analyze and visualize huge volumes of data generated by our digital and computing world. Several challenges in handling petabytes of information, commonly named as Big data needs to be addressed in more efficient way. Big data management (BDM) is the process of collecting, storing, analysing and visualization of large volumes of data, which can be in the form of structured, unstructured and semi-structured formats. Problems such as data acquisition, data storage, data retrieval, data analysis, and data visualization can no longer be handled by traditional database systems. The primary purpose of this paper is to provide a comprehensive survey on Big data management and to provide an overview on various algorithms related to job scheduling in Hadoop and the latest advancements. These research directions can lead to exploration of Big data domain and result in development of optimal techniques and scheduling algorithms to address problems faced in Big data.
References


19. Jiong Xie, Shu Yin, Xiaojun Ruan, Zhiyang Ding, Yun Tian, James Majors, Adam Manzanares and Xiao Qin, “Improving MapReduce Performance through Data Placement in
Heterogeneous Hadoop Clusters,” in Proc., 19th International Heterogeneity in Computing Workshop, Atlanta, Georgia, 2010.


43. Xie Jiong, Yin Shu, Ruan Xiaojun, Ding Zhiyang, Tian Yun, “Improving Mapreduce performance through data placements in heterogeneous hadoop cluster,” 2010.

Index Terms

Computer Science           Distributed Computing

Keywords

Big data, Big data management, Job Scheduling, Hadoop, MapReduce.