Abstract

In today's world the amount of data being generated is growing exponentially and use of internet is also increasing it leads to handle lots of data by internet service providers. MapReduce is one of the good solutions for implementing large scale distributed data application. A MapReduce workload generally contains a set of jobs, each of job consists of multiple map and reduce tasks. Map task executed before reduce task and map tasks can only run in map slot and reduce tasks can only run in reduce slot. Due to that different job executions orders and map/reduce slot configurations for a MapReduce workload have different performance metrics and different system utilization. Makespan and total completion time are two key performance metrics. This paper proposes two algorithm for these two key metrics, The first class of algorithms mainly focuses on the job ordering optimization for a MapReduce workload under given slot configuration and the second class of algorithms perform optimization for slot configuration for a MapReduce workload.

Index Terms

Computer Science

Distributed Computing

Keywords

MapReduce, Hadoop, Flow-shops, Scheduling algorithm, Job ordering.