Abstract

MapReduce implementations are being used for processing large data sets. MapReduce performs parallel computations to speed up the job processing. When performing parallel computations the skew that arises due large indivisible records or uneven distribution of data slows down the job execution process and lowers the cluster throughput. We provide a solution, by proposing an automatic system that handles skew which is compatible with MapReduce framework and is transparent to users. The proposed system makes use of idle resources in the cluster for skew handing. Task repartitioning method is implemented for the purpose of skew handling. The output order is maintained even after task repartitioning. The proposed system requires no extra input from the users and imposes minimum overhead in the absence of skew.

References


Index Terms

Computer Science

Information Sciences
Keywords
Data skew  MapReduce  parallel database systems  performance gain  skew handling