Hash Semi Cascade Join for Joining Multi-Way Map Reduce

مكان النشر

IEEE SAI Intelligent Systems Conference 2015 November 10-11, 2015 | London, UK

أسماء المشتركين فى البحث :

Marwa Hussien Mohamed- Mohamed Helmy Khafagy

ملخص البحث باللغة الانجليزية

Map-reduce is a programming model popularized by Google since 2004. It's used with large-scale datasets and processing data on a shared-nothing cluster. Map-Reduce accomplish high performance by partitioning the processes into small units of work that can run in parallel across thousands of nodes in the cluster. Rapidly, increasing in data size has rise importance to uncover hidden pattern to acquire new knowledge and get valuable information. But, map-reduce doesn't directly support join operation. This paper discusses some types of two-way algorithms, list some advantage and disadvantage of every algorithms. We propose a new multi - way join algorithm hash semi cascade join used to join more than two data sets. Using hash tables in the first phase, deleting unused records for joint operation as early as possible to reduce network bottleneck and increase performance. We compare this new algorithm with some types of multi-way join like map side join, reduce side one shot join and reduce side cascade join. Our results show that the map side join has more time for sorting data and do join result with small data sets with high performance but, time increase while data are increased. Reduce side one shot join has join result near map side join. Reduce side cascade join get more time to get the final result. Hash semi cascade join gain high performance using hash tables. According to, reduce shuffling records as in reduce side one shot and reduce side cascade join it can do join for any data set size.