Abstract

The elastic scalability and fault tolerance of the cloud computing has led to a wide range of real world applications. However, processing requirements of Big Data in these applications pose a humongous challenge for achieving desired performance levels. MapReduce is an effective parallel distributed programming model for handling large unstructured datasets in cloud applications. Hadoop, an open source implementation of the MapReduce model, is currently being employed for high performance processing of Big Data. The current Hadoop implementation considers the nodes of a cluster in a homogeneous environment where each node has the same computing capacity and workload. But in real world applications the nodes
may have different computing capacities and workloads resulting in a heterogeneous environment. In such heterogeneous environment the default Hadoop implementation does not yield the expected performance. This paper includes a survey on the algorithms proposed by different authors on (a) data placement strategies and (b) workload scheduling for Hadoop in heterogeneous network.

References

- Bin Ye, Xiaoshe Dong, Pengfei Zheng, Zengdong Zhu, Qiang Liu, Zhe Wang, "A
Delay Scheduling Algorithm based on History Time in Heterogeneous Environments,


- Jiong Xie, Shu Yin, Xiaojun Ruan, Zhiyang Ding, Yun Tian, James Majors, Adam Manzanares, Xiao Qin, Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters, in proceedings of International Symposium on Parallel and Distributed Processing, Workshops and PhD Forum, pp. 1-9 Apr. 2010.


Index Terms

Computer Science  Circuits And Systems
Keywords
Cloud Computing  Big Data  Mapreduce  Hadoop  Heterogeneous Network.