Abstract

The majority of large-scale data severe applications executed by data centers are based on MapReduce or its open-source implementation i.e. Hadoop. For processing huge sum of data in parallel Hadoop programming framework provides Distributed File System (HDFS)[2] and MapReduce Programming Model[3]. Job scheduling is an imperative process in Hadoop MapReduce. Hadoop comes with three types of schedulers namely FIFO, Fair and Capacity Scheduler. In some processing scenario these traditional scheduling algorithm of Hadoop cannot meet the performance requirements and fairness criteria of Big Data Processing. To address this issue new efficient scheduler is require who can identify the data size first and processed accordingly for performance improvement. This new MapReduce scheduling
scheme Will improves MapReduce performance and erasure high speed data processing. Proposed system will analyze the data size of individual DataNode and create threads based on threshold value decided by proposed scheduler. Processing of the threads is done parallel on individual DataNode by task tracker which will ultimately improve the data process performance. Because of that task Tracker will does the work in less time than the time required by the traditional Scheduler.

References

- Apache Hadoop. Available at http://hadoop.apache.org
- ApacheHDFS. Available at http://hadoop.apache.org/hdfs

- A community white paper developed by leading researchers across the United States &quot;Challenges and Opportunities with Big Data&quot;
- Jeffrey Dean and Sanjay Google, Inc. &quot;MapReduce: Simplified Data Processing on Large Clusters&quot;
- KyuseokShimSeoulNationalUniversityshim@ee.snu.ac.kr &quot;MapReduce Algorithms for Big Data Analysis&quot;
- Vasiliki Kalavri, Vladimir VlassovKTH The Royal Institute of Technology Stockholm, Sweden kalavri@kth.se &quot;MapReduce: Limitations, Optimizations and Open Issues&quot;
- TrustCom/ISPA/IUCC,Page1031-1038,IEEE,(2013)
- Yi Yao, Jianzhe Tai, Bo Sheng, Ningfang Mi, &quot;LsPS: A Job Size-Based Scheduler for Efficient Task Assignments in Hadoop&quot; In proceedings of the IEEE transaction, Copyright (c) 2014 IEEE
- Qutaibah Althebyan, Omar ALQudah, Yaser Jararweh Qussai Yaseen &quot;Multi-Threading Based Map Reduce Tasks Scheduling&quot; 2014 5th International Conference on Information and Communication Systems (ICICS)
- Jisha S Manjaly, Varghese S Choorallil Department &quot;TaskTracker Aware Scheduling for Hadoop MapReduce&quot; 2014 5th International Conference on Information and Communication Systems (ICICS)
- Runhui Li, Patrick P. C. Lee, Yuchong Hu &quot;Degraded-First Scheduling for MapReduce in Erasure-Coded Storage Clusters&quot; AoE/E-02/08 and ECS CUHK419212 from the University Grants Committee of Hong Kong, IEEE 2013
- Bin Ye, Xiaoshe Dong, Pengfei Zheng &quot;A delay scheduling algorithm based on history time in heterogeneous environments&quot; 2013 8th Annual ChinaGrid Conference
- S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, &quot;An analysisof traces from a
production mapreduce cluster; in CCGRID'10, 2010, pp. 94–103.

Index Terms

Computer Science

Information Sciences

Keywords

Mapreduce  Big Data  Scheduling  Hdfs