International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 4, July 2014, 1-10

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

AN OPTIMAL DATA PLACEMENT FRAMEWORK TO INCREASE PERFORMANCE BOF MAPREDUCE FOR DATA-INTENSIVE APPLICATIONS WITH INTEREST LOCALITY

SOLMAZ VAGHRI[1] & MOHAN K. G[2]

[1]M.Tech Student, Department of CSE, Acharya Institute of Technology, Affiliated to VTU, Bangalore, Karnataka, India

[2]Professor, Department of CSE, Acharya Institute of Technology, Affiliated to VTU, Bangalore, Karnataka, India

ABSTRACT

Emerging many numbers of data-intensive applications that needs to access ever-increasing data sets ranging from gigabytes to terabytes or even petabytes, place a demand on employing parallel processing techniques to optimize performance and reduce the decision time. Of late parallel computing frameworks such as MapReduce and it's open source implementation Apache Hadoop has been used to run large scaledata-intensive applications and conduct analysis, but data locality have not been taken into account in Hadoop and MapReduce and they use random data distribution method for load balancing. Practically in many data-intensive applications data groups often accessed to gather and only subset of a whole data set are frequently used. Ignoring data grouping issue and random data placement noticeably reduce the performance of MapReduce and Hadoop. This paper presents architecture and implementation status of a an optimal data placement framework that dynamically analyzes data accesses from system log files and create optimal data groupings and distribute the data evenly to achieve maximum parallelism per data group and significantly improves the overall performance of MapReduce for data-intensive applications.

KEYWORDS: Data-Intensive, Data Placement, Hadoop, Map Reduce, Parallel Processing