Ibrahim Abaker Targio Hashem et al., J Data Mining In Genomics & Proteomics 2015, 6:4
http://dx.doi.org/10.4172/2153-0602.C1.003

OMICS International
conferenceseries.com

2nd International Conference on
Big Data Analysis and Data Mining
November 30-December 01, 2015 San Antonio, USA

Schedule optimization for big data processing on cloud

Ibrahim Abaker Targio Hashem, Nor Badrul Anuar and **Abdullah Gani**
University of Malaya, Malaysia

Over the past few years, the continuous increase in computational capacity has produced an overwhelming flow of data or big data, which exceeds the capabilities of conventional processing tools. Big data offer a new era in data exploration and utilization. The major enabler for underlying many big data platforms is certainly the MapReduce computational paradigm. MapReduce is recognized as a popular programming model for the distributed and scalable processing of big data and is increasingly being used in different applications mostly because of its important features that include scalability, flexibility, ease of programming, and fault-tolerance. Scheduling tasks in MapReduce across multiple nodes have shown to be multi-objective optimization problem. The problem is even more complex by using virtualized clusters in a cloud computing to execute a large number of tasks. The complexity lies in achieving multiple objectives that may be of conflicting nature. For instance, scheduled tasks may require to make several tradeoffs between the job performance, data locality, fairness, resource utilization, network congestion and reliability. These conflicting requirements and goals are challenging to optimize due to the difficulty of predicting a new incoming job's behavior and its completion time. To address this complication, we introduce a multi-objective approach using genetic algorithms. The goal is to minimize two objectives: Execution time, and budget of each node executing the task in the cloud. The contribution of this research is to propose a novel adaptive model to communicate with the task scheduler of resource management. The proposed model periodically queries for resource consumption data and uses to calculate how the resources should be allocated to each task. It passes the information to the task scheduler by adjusting task assignments to task nodes accordingly. The model evaluation is realized in scheduling load simulator. PingER, the Internet End-to-End performance measurement, was chosen for performance analysis of the model. We believe this proposed solution is timely and innovative as it provides a robust resource management where users can perform better scheduling for big data processing in a seamless manner.

Biography

Ibrahim Abaker Targio Hashem is currently a PhD degree candidate at the Department of Compute Systems, UM. He has been working on Big Data since 2013 and his article on Big Data becomes top most downloaded in 2014 *Information System* journal of Elsevier. He has experience in configuring Hadoop MapReduce in multi-node cluster. His main research interests include big data, cloud computing, distributed computing, and network.

targio@siswa.um.edu.my

Notes: