ROYAL UNIVERSITY OF PHNOM PENH

Master of Science in Information
Technology Engineering

Software Development

PERFORMANCE EVALUATION OF TEXT PROCESSING USING APACHE HADOOP FRAMEWORK

Advisors: Mr. Taing Nguonly

Keywords: Map/Reduce, Hadoop, Distributed File System

Field related: Distributed Systems, Text Processing, Advance Parallel Programming Language

Abstract

Text processing, for example word segmentation, requires a lot of power to run. It can take days or weeks to execute large scale of corpus on a single machine. Since Google released Map/Reduce framework, Apache group formed a team to establish Hadoop for free distribution under GPL licensed. Apache Hadoop is a framework for running application on a large cluster built of commodity hardware. Hadoop takes advantage of software paradigm Map/Reduce framework plus Hadoop Distributed File System (HDFS) to leverage the significant performance for text processing. This article tends to seek for understanding of the Hadoop architecture, Distributed File System and Map/Reduce framework. And finally it presents the performance evaluation of text processing based on Hadoop framework.

References:

1. GHEMAWAT, S., GOBIOFF, H., AND LEUNG, S.-T. The Google File System. In Proc. of the 19th ACM SOSP (Dec. 2003), pp. 29⍰43
2. http://hadoop.apache.org/, 2011