Large-Scale Web Traffic Log Analyzer using Cloudera Impala on Hadoop Distributed File System

Choopan Rattanapoka* and Prasertsak Tiawongsombat

Abstract

Resource planning and data analysis are important for network services in order to increase the service efficiency. Nowadays, Large websites or web servers have a large number of visitors, which mean a large web traffic log need to be stored in the plain text or the relational database. However plain text and relational database are not efficient to handle a large number of data. Moreover, the web traffic log analysis hardware or software that can handle such a big data is also expensive. This research paper proposes the design of a large-scale web traffic log analyzer using PHP language to show the visitors' traffic data analysis in the form of charts. The Hadoop Distributed File System (HDFS) is used in conjunction with other related techniques to gather and store visitors' traffic log. Cloudera Impala is used to query web traffic log stored in HDFS while Apache Thrift is an intermediary connecting Cloudera Impala to PHP web. Upon testing our large-scale web traffic log analyzer on HDFS Cluster of 8 nodes with 50 gigabytes of traffic log, our system can query and analysis web traffic log then display the result in about 4 seconds

Keywords: Hadoop Distributed File System, Cloudera Impala, Big Data, Log Analysis, PHP

Department of Electronics Engineering Technology, College of Industrial Technology, King Mongkut University of Technology North Bangkok.

* Corresponding author, E-mail: choopan.r@cit.kmutnb.ac.th Received 30 March 2016, Accepted 16 December 2016