Content:
Apache Hadoop is a software framework which is capable of distributed processing of large datasets across clusters of machines using MapReduce programming model. Hadoop system has supported a wide variety of applications including text analysis, business model and social network analysis. With the explosively growing of the scale of data and clusters, system optimization is a more important issue for large scale systems because of the expensive cost of maintaining the system.

In this paper, we focus on the configuration parameters tuning, which is the most common way of system optimization. However, two challenges exist: (1) It is hard for administrators to specify parameters correctly because the behavior of such a large scale system is too complex. (2) Dozens of configuration parameters are involved to the performance which also makes the configuration parameter tuning more difficult.

We propose an analytical method to help the system administrator to choose an approximately optimal configuration parameters depending on the characteristics of each application. Our approach has two key phases: prediction and optimization. The prediction is to estimate the performance of a Hadoop MapReduce job whereas the optimization is to search the approximately optimal configuration parameters strategically by invoking the prediction part repeatedly. (1) The prediction part is composed of several sub-model built by machine learning algorithm. The structure of prediction is according to how the dataflow is operated through each phase of Hadoop MapReduce. (2) In conceptual manner, the optimization part is RRS (recursive random search), which is the extended search method of random sampling. Optimization is based on prediction part as the objective function to search the approximately optimal configuration parameters. Besides the configuration parameter tuning, we also show the relevance of parameters to performance and talk about deciding what parameters should be tuning for different kind of applications.

Primary authors: Mr. CHEN, Chi-Ou (Industrial Technology Research Institute of Taiwan, R.O.C)
Co-authors:
Presenter: Mr. CHEN, Chi-Ou (Industrial Technology Research Institute of Taiwan, R.O.C)
**Session classification**: Big Data II

**Track classification**: Big Data

**Type**: Oral