I/O Performance Analysis for Big Data Clustering through Compression Contribution Model

Yang Zhang, Jianhui Li, Yuanchun Zhou, Zhenghua Xue and Geng Shen,
Computer Network Information Center, Chinese Academy of Sciences, Beijing

In the era of big data analysis, I/O becomes an inevitable challenge. Compression technology can sigfnificantly alleviate the increasingly prominent I/O bottleneck occured in big data analysis. This paper, taking parallel K-means clustering as an example, proposed a compression contribution model. We investigates multiple factors related to compression in depth, including the compression ratio, the number of computing cores, the compression/decompression speed, the data size and when and how to use compression and so on.

Based on experiments on 1.12 TeraBytes data via computing clusters with 336 computing cores, the measurement result shows that integrating the compression into big data analyais yields significant performance improvement. Compression contribution model is capable of providing effective decision support for when and how to use the compression to improve I/O performance.

Keywords: Big Data Analysis, I/O Bottleneck, Compression Contribution Model