Analyzing and Optimizing the Performance of Big Data Platform: A Case Study Based on Apache Hadoop MapReduce Framework

J.H.R.P. Jayamaha* and K.P.N. Jayasena

Department of Computing and Information Systems, Faculty of Applied Sciences,
Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.

*rpjayamaha@gmail.com

Map-reduce is among the most effective and efficient methods to handle many data sets. Different methods and techniques have been presented to map-reduce processes. Large-scale data processing and analysis can be performed using Apache Hadoop distributed framework on commodity equipment. Parameters can be tweaked in Hadoop, and they have a significant impact on the performance of MapReduce applications. Hadoop set-up parameter adjustment is an excellent way to boost the performance. New research areas have emerged based on the Hadoop map-reduce framework. Performance optimization is mainly based on different concurrent containers and a suitable Hadoop Distributed File System (HDFS). When considering concurrent containers, it is based on CPU performance, network parameters, and memory utilization. All those factors impact the performance of Hadoop map-reduce framework. In this study, we consider the above factors in optimizing the performance of the Apache Hadoop MapReduce framework. In this study, we optimize container performance and Hadoop HDFS block. The primary outcome of this project is to introduce the best system architecture and suitable Hadoop HDFS block size. This performance tuning is the most advantageous process in Apache Hadoop. In this experiment, we analyzed the default Hadoop map-reduce process performance. After the performance optimization in the Hadoop framework, this system implementation significantly improves the Bigdata Map reducing process. According to the experiment, HDFS block size depends on the Hadoop MapReduce performance. If the dataset grows larger, the HDFS block size must be increased to improve performance. Also, the concurrent container performance may highly affect the performance of the process. Also, concurrent container memory size is more effective rather than the CPU count. All of these factors were determined after multiple trials to yield accurate results. All of these factors have a significant impact on the performance of Hadoop MapReduce.

Keywords: Apache Hadoop, Concurrent Container, Hadoop Distributed File System, Map-Reduce