World Academy of Science, Engineering and Technology
International Journal of Urban and Civil Engineering
Vol:11, No:05, 2017

Regression Approach for Optimal Purchase of Hosts Cluster in Fixed Fund for Hadoop Big Data Platform

Authors : Haitao Yang, Jianming Lv, Fei Xu, Xintong Wang, Yilin Huang, Lanting Xia, Xuewu Zhu

Abstract : Given a fixed fund, purchasing fewer hosts of higher capability or inversely more of lower capability is a must-be-made trade-off in practices for building a Hadoop big data platform. An exploratory study is presented for a Housing Big Data Platform project (HBDP), where typical big data computing is with SQL queries of aggregate, join, and space-time condition selections executed upon massive data from more than 10 million housing units. In HBDP, an empirical formula was introduced to predict the performance of host clusters potential for the intended typical big data computing, and it was shaped via a regression approach. With this empirical formula, it is easy to suggest an optimal cluster configuration. The investigation was based on a typical Hadoop computing ecosystem HDFS+Hive+Spark. A proper metric was raised to measure the performance of Hadoop clusters in HBDP, which was tested and compared with its predicted counterpart, on executing three kinds of typical SQL query tasks. Tests were conducted with respect to factors of CPU benchmark, memory size, virtual host division, and the number of element physical host in cluster. The research has been applied to practical cluster procurement for housing big data computing.

Keywords : Hadoop platform planning, optimal cluster scheme at fixed-fund, performance predicting formula, typical SQL query tasks
Conference Title : ICSCTB 2017 : International Conference on Smart City, Transportation and Buildings
Conference Location : London, United Kingdom
Conference Dates : May 25-26, 2017