World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:14, No:10, 2020

Time and Cost Prediction Models for Language Classification Over a Large Corpus on Spark

Authors : Jairson Barbosa Rodrigues, Paulo Romero Martins Maciel, Germano Crispim Vasconcelos

Abstract : This paper presents an investigation of the performance impacts regarding the variation of five factors (input data size, node number, cores, memory, and disks) when applying a distributed implementation of Naïve Bayes for text classification of a large Corpus on the Spark big data processing framework. Problem: The algorithm's performance depends on multiple factors, and knowing before-hand the effects of each factor becomes especially critical as hardware is priced by time slice in cloud environments. Objectives: To explain the functional relationship between factors and performance and to develop linear predictor models for time and cost. Methods: the solid statistical principles of Design of Experiments (DoE), particularly the randomized two-level fractional factorial design with replications. This research involved 48 real clusters with different hardware arrangements. The metrics were analyzed using linear models for screening, ranking, and measurement of each factor's impact. Results: Our findings include prediction models and show some non-intuitive results about the small influence of cores and the neutrality of memory and disks on total execution time, and the non-significant impact of data input scale on costs, although notably impacts the execution time.

Keywords : big data, design of experiments, distributed machine learning, natural language processing, spark
Conference Title : ICDCS 2020 : International Conference on Distributed Computer Systems
Conference Location : Beijing, China
Conference Dates : October 07-08, 2020