World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:9, No:09, 2015

Big Data Analysis with Rhipe

Authors : Byung Ho Jung, Ji Eun Shin, Dong Hoon Lim

Abstract : Rhipe that integrates R and Hadoop environment made it possible to process and analyze massive amounts of data using a distributed processing environment. In this paper, we implemented multiple regression analysis using Rhipe with various data sizes of actual data. Experimental results for comparing the performance of our Rhipe with stats and biglm packages available on bigmemory, showed that our Rhipe was more fast than other packages owing to paralleling processing with increasing the number of map tasks as the size of data increases. We also compared the computing speeds of pseudo-distributed and fully-distributed modes for configuring Hadoop cluster. The results showed that fully-distributed mode was faster than pseudo-distributed mode, and computing speeds of fully-distributed mode were faster as the number of data nodes increases.

Keywords : big data, Hadoop, Parallel regression analysis, R, Rhipe
Conference Title : ICASMA 2015 : International Conference on Applied Simulation, Modelling and Analysis
Conference Location : Berlin, Germany
Conference Dates : September 14-15, 2015