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

Big Data Analysis with RHadoop

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

Abstract : It is almost impossible to store or analyze big data increasing exponentially with traditional technologies. Hadoop is a new technology to make that possible. R programming language is by far the most popular statistical tool for big data analysis based on distributed processing with Hadoop technology. With RHadoop that integrates R and Hadoop environment, we implemented parallel multiple regression analysis with different sizes of actual data. Experimental results showed our RHadoop system was much faster as the number of data nodes increases. We also compared the performance of our RHadoop with lm function and big lm packages available on big memory. The results showed that our RHadoop was faster than other packages owing to paralleling processing with increasing the number of map tasks as the size of data increases.

Keywords : big data, Hadoop, parallel regression analysis, R, RHadoop

Conference Title : ICASMA 2015 : International Conference on Applied Simulation, Modelling and Analysis

Conference Location : Berlin, Germany

Conference Dates : September 14-15, 2015

Open Science Index, Computer and Systems Engineering Vol:9, No:09, 2015 publications.waset.org/abstracts/36106.pdf