Open Science Index, Mathematical and Computational Sciences Vol:14, No:12, 2020 publications.waset.org/abstracts/41570.pdf

Improved K-Means Clustering Algorithm Using RHadoop with Combiner

Authors : Ji Eun Shin, Dong Hoon Lim

Abstract : Data clustering is a common technique used in data analysis and is used in many applications, such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing. K-means clustering is a well-known clustering algorithm aiming to cluster a set of data points to a predefined number of clusters. In this paper, we implement K-means algorithm based on MapReduce framework with RHadoop to make the clustering method applicable to large scale data. RHadoop is a collection of R packages that allow users to manage and analyze data with Hadoop. The main idea is to introduce a combiner as a function of our map output to decrease the amount of data needed to be processed by reducers. The experimental results demonstrated that K-means algorithm using RHadoop can scale well and efficiently process large data sets on commodity hardware. We also showed that our K-means algorithm using RHadoop with combiner was faster than regular algorithm without combiner as the size of data set increases.

Keywords : big data, combiner, K-means clustering, RHadoop
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020