European Journal of Scientific Research
ISSN 1450-216X / 1450-202X Vol. 141 No 4 Setember, 2016, pp.374-383
http://www.europeanjournalofscientificresearch.com

A Framework for Distributed Data Mining of Big Data through Machine Learning

Saif Khalid Musluh
Assistant Lecture Specialization Computer Science Information System
College of Biotechnology, University of AL-Qadisiyah
AL-Qadisiya – AL-Dewaniya - Iraq
E-mail: saif.khalid@qu.edu.iq

Abstract

Machine learning applications have been around for many years. However, machine learning approaches on big data is relatively new phenomena. Big data processing needs to be done in a distributed environment where thousands of nodes work together along with data centres, cloud computing resources and distributed file systems. Hadoop is one such example for distributed programming framework. The programming approach used in such environment is known as MapReduce which is best used to let multiple nodes participate in processing given job. It exploits the power of GPU and supports parallel processing of data. The existing machine learning algorithms focused processing huge amount of data. However, there needs to be a framework that can improve the support for big data processing using machine learning algorithms. In this paper we proposed and implemented a framework that can support information retrieval and natural language processing. The framework takes big data as input and provides simulated environment demonstrating the functionalities of mapper and reducer. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that the framework supports distributed data mining of big data through machine learning. However it is in its primitive stage and needs enhancement to be more useful in future.

Index terms: Big data, distributed programming framework, MapReduce, machine learning

Introduction

Big data refers to the data which is bulky and exhibits exponential growth. The data which has certain characteristics is known as big data. Those characteristics are known as Volume, Variety and Velocity. The first attribute is related to the quantity of data which is very huge in nature. The second attribute refers to the fact that data is available in different formats like structured, semi-structured and unstructured. The third attribute that is Velocity refers to the data that is in transit. In other words the data which is moving or streaming shows the third attribute velocity of data. When live data is being collected from different sources, such data can exhibit velocity characteristic. Big data when accumulated becomes a valuable source for extracting business intelligence. Unfortunately the process of big data cannot be directly done in the local machines due to resource constraints. It is possible with distributed programming frameworks that are associated with huge data centres and cloud computing paradigm.