5th International Conference of Multidisciplinary Approaches (iCMA), 2018
Faculty of Graduate Studies,
University of Sri Jayewardenepura,
Sri Lanka
ISSN: 2386 – 1509

APPLICATION OF THE MapReduce PROGRAMMING FRAMEWORK TO GENOME BIG DATA ANALYSIS

Lekamge T.C.* and Silva M.D.R.L.
Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Sri Lanka
thisuri.lekamge94@gmail.com

ABSTRACT

Today the world mainly associates with large amount of data sets. Especially in genome world, there are terabytes of data to store and analyze. Also there are data mining tools which need to be developed for petabytes of data. If we get, the volume of data use in 'Facebook', need to be collected and managed on a daily basis. Those kinds of data can fall under the category of big data. Big data is a collection of large datasets that cannot be processed using traditional computing techniques. This term; Big data is used to describe huge datasets which involves 4V definition; volume, variety, velocity and veracity. Therefore big data technology is important in providing more accurate analysis since it may lead to more concrete decision making giving greater operational efficiencies and cost and risk reduction. In that manner this is very important in genome analysis because there are lots of challenges in storing and analyzing genome data. These are called "Big data challenges". There are two classes of technology; operational big data and analytical big data. Analytical big data technology is used here. This is consisted with MapReduce and Massively Parallel Processing (MPP) database systems that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. In this paper, I review the existing applications of MapReduce programming framework and its implementation platform Hadoop and how they support in genome big data analysis. The objective of this paper is to summarize the state-of-art efforts in genome big data analytics and highlight what might be needed to enhance the outcomes of genome big data analytics tools. Materials and methodologies and research findings and results sections have been presented respectively. In discussion, we discuss some developments and limitations in tools. Final section presents the conclusion.

Keywords: Bigdata, Genome Analysis, MapReduce, Hadoop, Massively Parallel Processing (MPP)