ABSTRACT

Big Data is currently a hot topic among organizations and researchers around the world due to the emergence of new technologies and media. Hadoop Mapreduce and Apache Spark are two popular opensource frameworks for processing large amounts of data. The purpose of this research is to measure the level of accuracy and performance with the Support Vector Machine algorithm in the process of comparing Apache Spark and Hadoop Mapreduce. A multi-node cluster will be created where 3 VMs will be created, each will have Apache Spark and Apache Hadoop installed, with the rule that 1 VM will act as a master and the other two as workers. The datasets that already exist in Hadoop Mapreduce and Apache Spark are then tested using the SVM algorithm. The results, after 9 tests, found that there are 9 different accuracies where the average accuracy is at 80 % and then there is a difference in speed in training time and testing time where Apache Spark in this research is 3 times faster than Apache Hadoop.

Keywords : Big Data, Hadoop Mapreduce, Apache Spark, Support Vector Machine.

ABSTRAK

Big Data saat ini menjadi topik hangat di kalangan organisasi dan peneliti seluruh dunia karena munculnya teknologi dan media baru. *Hadoop Mapreduce* dan *Apache Spark* adalah dua *framework opensource* populer untuk memproses data dalam jumlah besar. Tujuan dalam penilitian ini yaitu Mengukur tingkat akurasi dan performa dengan algoritma *Support Vector Machine* pada proses perbandingan *Apache Spark* dan *Hadoop Mapreduce*. *Cluster multi-node* akan dibuat dimana 3 *VM* akan dibuat, masing-masing akan menginstal *Apache Spark* dan *Apache Hadoop*, dengan aturan bahwa 1 *VM* akan bertindak sebagai master dan dua lainnya sebagai pekerja. Dataset yang sudah ada dalam *Hadoop Mapreduce* dan *Apache Spark* lalu diuji menggunakan algorirtma *SVM*. Hasilnya, setelah dilakukan 9 pengujian menemukan bahwa terdapat 9 akurasi yang berbeda dimana rata-rata akurasinya ada pada nilai 80% lalu terdapat perbedaan kecepatan pada training time dan testing time dimana *Apache Spark* dalam penilitian ini lebih cepat 3 kali lipat dibandingkan dengan *Apache Hadoop*.

Kata kunci : *Big Data, Hadoop Mapreduce, Apache Spark, Support Vector Machine.*