ABSTRAK

Perkembangan data yang sangat pesat membuat teknologi *big data* menjadi inovasi baru dalam menyimpan data. Apache Hadoop merupakan framework *big data* yang mampu menyimpan data tanpa memperhatikan jenis data. Apache Hadoop menggunakan model pemrograman MapReduce dalam menganalisa data. Apache Mahout merupakan *library* analisa data yang mampu menjalankan komputasi berbasis pemrograman MapReduce. Apache Mahout telah menyediakan komputasi penambangan data yang dapat digunakan dalam menganalisa data. K-Means merupakan metode penambangan data yang dapat mengelompokkan data berdasarkan kemiripan sifat.

Penelitian ini menggunakan 4 komputer klaster yang berjalan pada jaringan lokal. Apache Hadoop yang berjalan pada sistem Linux dibagi menjadi 1 *master slave* dan 3 *slave node*. *Master node* mengatur komputasi MapReduce. *Slave node* bertugas sebagai media penyimpan data. Hasil K-Means dengan menggunakan *library* Mahout diuji dengan hasil dari metode manual. Hasil pengujian menunjukkan bahwa *library* Mahout mampu memberikan hasil analisa dengan benar. Sedangkan pengujian unjuk kerja dilakukan dengan menjalankan K-Means sebanyak 10 kali pada jumlah *slave node* yang berbeda. Kesimpulan unjuk kerja sistem Hadoop dilakukan dengan mencari nilai rata-rata dari percobaan-percobaan tersebut. Hasil unjuk kerja menunjukkan bahwa semakin banyak jumlah *slave node* maka semakin cepat proses komputasi.

Kata Kunci: *Big Data*, Hadoop, MapReduce, Mahout, *Data Mining*, K-Means

ABSTRACT

The growth of massive data makes big data technology as a new innovation in storing data. Apache Hadoop is a big data framework that able to stroing data without considering the variety of data. Apache Hadoop uses MapReduce programming model to analyze data. Apache Mahout is a data analyze library that able to analyze data in MapReduce programming model. Apache Mahout has provided data mining method as analyze data algorithm. K-Means is a data mining algorithm that can group item data into specific cluster based on similarity measure.

This research is developed in 4 computer cluster which is clustered in local network. Apache Hadoop that is adopted in Linux system is divided into 1 master node and 3 slave nodes. Master node handles MapReduce. Slave nodes roles as storage system. The output of K-Means Mahout library is evaluated with manual calculation. The evaluation result describe that Mahout library can analyze data well. The performance of Hadoop system is evaluated by running 10 times of K-Means with Mahout library in difference quantity of slave node. The conclusion is taken by calculate the mean value of each 10 trainings. The performance evaluation result explained that increasing the number of slave node can make time execution of computation to be faster.

Keyword : Big Data, Hadoop, MapReduce, Mahout, Data Mining, K-Means