INTISARI

PENERAPAN METODE *CLUSTERING* PADA *SPARK* DENGAN STUDI KASUS DATA BUS RAPID TRANSIT

Oleh

Andhika Kurnia Harryajie

13/348657/PA/15465

Dunia teknologi semakin cepat berkembang di era modern ini. Salah satu permasalahan yang dihadapi pada sistem *Bus Rapid Transit (BRT)* adalah pengolahan data penumpang yang terus bertambah sesuai dengan pertumbuhan jumlah dan tingkat keramaian penduduk di suatu kota. Data yang semakin bertambah akan sulit dikelola dengan sistem *Relational Database Management System (RDBMS)* karena pada proses pengolahan data membutuhkan waktu yang lebih lama dan akan semakin berkurang efisiensinya. Sehingga pemrosesan paralel dapat menjadi sebuah solusi dalam mengolah data serta meningkatkan waktu pemrosesan yang lebih cepat. *Framework Hadoop* digunakan untuk mendukung *environment* pemrosesan paralel sedangkan untuk proses menganalisis data digunakan *platform Spark*. Kelebihan yang dimiliki dari *Spark* yaitu memiliki kemampuan proses yang lebih cepat karena memanfaatkan proses di *memory* dan sering digunakan untuk analisis *BigData*.

Analisis *BigData* yang dilakukan pada penelitian ini yaitu analisis data dengan menggunakan metode *clustering* sebagai tahapan dalam *data mining*. Metode *clustering* merupakan teknik pengambilan data atau informasi yang mirip di suatu kelompok atau klaster dan data yang tidak sama dengan kelompok lainnya.

Pengujian *Sum of Square Error* dan *Silhouette coefficient* dilakukan pada dua algoritme *K-Means* dan *Scalable K-Means++* dalam melakukan pengelompokan jumlah penumpang terhadap waktu tertentu. Hasil pengujian *Sum of Square Error* dan *Silhouette coefficient* menghasilkan *k* klaster yang optimal yaitu pada jumlah 8 klaster. Kemudian untuk pengujian *Silhouette coefficient* pada *Scalable K-Means++* lebih besar yaitu 0.57380692 dibandingkan dengan hasil *Silhouette coefficient* dari *K-Means* yaitu 0.57105644. Sehingga dari hasil tersebut dapat disimpulkan bahwa algoritme Scalable K-Means lebih baik dibandingkan algoritme K-Means.

Kata kunci: *Clustering*, *Bus Rapid Transit*, *BigData*, *Spark*

ABSTRACT

IMPLEMENTATION OF *CLUSTERING* METHOD ON *SPARK* WITH CASE STUDY OF BUS RAPID TRANSIT DATA

By

Andhika Kurnia Harryajie

13/348657/PA/15465

The world of technology is growing rapidly in this modern era. One of the problems encountered in the Bus Rapid Transit (BRT) system is the growing data processing of passengers in accordance with the growing number and the level of crowds in a city. Increasing data will be difficult to manage with Relational Database Management System (RDBMS) system because data processing takes longer time and will decrease its efficiency. So parallel processing can be a solution in processing the data and increasing the processing time faster. The Hadoop framework is used to support parallel processing environments while for analyzing data Spark platforms are used. The advantages possessed of Spark is that it has a faster processing capability because it utilizes the process in memory and is often used for BigData analysis.

BigData analysis conducted in this research is data analysis by using clustering method as stages in data mining. Clustering method is a technique of data retrieval or similar information in a group or cluster and data that is not the same as other groups.

The Sum of Square Error and Silhouette coefficient evaluation were performed on two K-Means and Scalable K-Means ++ algorithms in grouping the number of passengers over a certain time. Result of testing of Sum of Square Error and Silhouette coefficient produce optimal cluster k which is at number 8 cluster. Then for testing of Silhouette coefficient on Scalable K-Means ++ is bigger that is 0.57380692 compared with result of Silhouette coefficient from K-Means that is 0.57105644. So from the results can be concluded that the algorithm Scalable K-Means better than K-Means algorithm.

*Key*words: *Clustering*, *Bus Rapid Transit*, *BigData*, *Spark*