Data integration simulation using data consolidation./ Hadaiq R. Sanabila, Ito Wasito

Hadaiq Rolis Sanabila, author

Abstrak

One of the data integration methods is data consolidation. This method captures data from multiple source systems/data and integrates it into a single persistent data. We examined the performance of data consolidation using k-means and Gaussian mixture clustering. Meanwhile, we use Silhouette index as cluster validation and measure how well of a clustering relative to others. The experiments analyses the data in various data duplication rate and actual number of data cluster. Based on the experimental result, there are two factors affecting the performance of data consolidation. These factors are the rate/percentage of duplicate data and the number of actual cluster contained in a data. The higher percentages of duplicate data and less number of clusters contained in the data would be increasing the performance of clustering algorithm.

Salah satu metode dari integrasi data adalah konsolidasi data. Metode ini mengambil data dari beberapa sumber data untuk digabungkan menjadi data persisten tunggal. Peneliti memeriksa kinerja konsolidasi data menggunakan beberapa teknik clustering yaitu k-means dan gaussian mixture clustering. Penulis menggunakan Silhouette index sebagai metode validasi cluster untuk mengukur seberapa baik suatu pengelompokan relatif terhadap data lain. Penelitian ini melakukan analisis data terhadap jumlah rata-rata duplikasi data dan jumlah sebenarnya dari cluster data. Berdasarkan hasil percobaan, ada dua faktor yang mempengaruhi kinerja integrasi data dengan menggunakan konsolidasi data. Faktor-faktor tersebut antara lain adalah tingkat atau persentase dari duplikasi data dan jumlah cluster sebenarnya yang terkandung dalam data. Persentase duplikasi data yang tinggi dan data yang mengandung jumlah cluster yang rendah, akan meningkatkan kinerja dari algoritma clustering.