Evaluasi Kecepatan Operasi dan Kepraktisan Penyimpanan Graf untuk Ver: View Discovery in the Wild = Operation Speed and Practicality Evaluation of Graph Storages for Ver: View Discovery in the Wild

Kevin Dharmawan, author

Abstrak

Ver adalah discovery system yang dibuat untuk mengidentifikasi join path pada data besar yang tidak mengandung join information. Ver menyimpan setiap kolom dari sumber data sebagai node dan potensi join path sebagai edge dalam bentuk graf menggunakan NetworkX. Namun, NetworkX memiliki limitasi pada besarnya graf yang dapat disimpan karena NetworkX menyimpan graf pada memory. Oleh karena itu, dibutuhkan alternatif penyimpanan graf yang menyimpan graf dalam persistent disk sebagai pengganti NetworkX pada Ver. Pencarian penyimpanan graf alternatif dilakukan dengan membandingkan beberapa penyimpanan graf yang meliputi: ArangoDB, CubicWeb, DGraph, DuckDB, IndraDB, JanusGraph, Kuzu, NebulaGraph, Neo4j, OrientDB, SurrealDB, dan TypeDB. Perbandingan dilakukan menggunakan graf acak dan graf dari dataset. Graf acak yang digunakan memiliki node dengan jumlah 100, 200, 400, 800, dan 1600 dengan kepadatan edge 0.1 sampai 1.0 dengan kenaikan 0.1. Dataset yang digunakan untuk perbandingan adalah TPC-H, ChEMBL, dan AdventureWorks. Perbandingan dilakukan dengan metode kuantitatif berdasarkan kecepatan operasi pemuatan data, 2-hop neighborhood, dan path finding serta metode kualitatif untuk kepraktisan dengan menilai kemudahan instalasi server, kemudahan implementasi client, dan kelengkapan dokumentasi. Didapatkan bahwa Kuzu adalah penyimpanan graf yang paling sesuai untuk menjadi pengganti NetworkX pada Ver.

.....Ver is a discovery system developed to identify join path in big data that doesn’t contain any join information. Ver stores each column of the data source as nodes and potential join path as edges in a graph using NetworkX. However, NetworkX has a limitation on the size of graph that can be stored because NetworkX stores graphs in memory. Therefore, an alternative graph storage that stores graphs in persistent disk is needed as a substitute for NetworkX on Ver. The search for alternative graph storage was carried out by comparing several graph storages which include: ArangoDB, CubicWeb, DGraph, DuckDB, IndraDB, JanusGraph, Kuzu, NebulaGraph, Neo4j, OrientDB, SurrealDB, and TypeDB. Comparisons are performed using random graphs and graphs from datasets. The ran dom graph used has node count of 100, 200, 400, 800, and 1600 with edge density 0.1 to 1.0 in increments of 0.1. The dataset used for comparison are TPC-H, ChEMBL, and AdventureWorks. Comparisons were made using a quantitative method based on the oper ation speed of data loading, 2-hop neighborhood, and path finding as well as a qualitative method for practicality by assessing the ease of server installation, the ease of implemen tation of client, and the completeness of documentation. It was found that Kuzu is the most suitable graph storage to replace NetworkX on Ver.