



Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Gráf információs rendszerek összehasonlító teljesítménymérése

Készítette

Antal János Benjamin
Elekes Márton

Konzulens

Marton József Ernő
Szárnyas Gábor

Kivonat

Az elmúlt évtizedben sokféle különböző NoSQL technikát használó adatbázis-kezelő készült. Ezek egyik csoportja a gráfadatbázisoké, melyek lehetővé teszik az adatok gráf formában történő tárolását és lekérdezését. Ez az adatmodell gyakran jobban illeszkedik a sok összefüggést tartalmazó adatok tárolására, mint a relációs modell, és a tömörsége miatt gyakran képes jobb teljesítményt nyújtani. Mindezek ellenére, mivel a relációs adatbázisokat majdnem 50 éve fejlesztik és optimalizálják, jelenleg is nyitott kérdés, hogy szükség van-e specializált gráfadatbázisokra a gráf adatok hatékony feldolgozásához.

Gráflekérdezések megfogalmazására új lekérdező nyelvek jelentek meg, mint például az openCypher. Ezeken a nyelveken gyakran kényelmesebben fogalmazhatunk meg gráflekérdezéseket, mint az SQL-alapú nyelveken. Az üzleti adatok jelentős része azonban jelenleg is hagyományos relációs adatbázisban van tárolva, emiatt ezen adatokat át kell tölteni gráfadatbázisokba, amely éles adatbázisok esetén általában nem megoldható. Célunk, hogy egy olyan megoldást készítsünk, amelyben lehetséges a magas kifejezőerővel rendelkező gráflekérdező nyelveken megfogalmazott lekérdezéseket hatékony relációs lekérdezőmotorokon futtatni anélkül, hogy szükség lenne egyik rendszerből a másikba áttölteni az adatokat. Ennek érdekében olyan fordítót (transpilert) készítettünk, ami képes openCypher lekérdezéseket SQL-re fordítani.

Különböző adatbázis-kezelő rendszerek összehasonlításához elengedhetetlenek a teljesítménymérési specifikációk (benchmarkok). Relációs adatbázisok esetében ezt a szerepet a Transaction Processing Performance Council benchmarkjai töltik be. A gráfadatbázisok relatív kiforrotlansága miatt jelenleg kevés benchmark létezik a gráflekérdezések teljesítménymérésére. Mi az LDBC (Linked Data Benchmark Council) Social Network Benchmark fejlesztésébe kapcsolódtunk be, amelynek keretében frissítettük és jelentősen fejlesztettük a meglévő implementációkat, továbbá elkészítettük a SPARQL nyelvű implementációt. Ezek felhasználásával alaposan megvizsgáltuk és részletesen elemeztük az adatbázis kezelőket különböző adatmodellek (relációs, gráf és szemantikus) felhasználásával.

Abstract

In the last decade, a lot of database management systems were developed with different NoSQL techniques. One group of these systems are graph databases, which allow users to store and query their data as graphs. This data model is often a better fit to represent strongly interlinked data sets than the traditional relational model, and its conciseness can lead to better performance. That said, relational databases have been developed and optimized for almost 50 years, and it is an open question whether efficient processing of graph data requires specialized databases.

New query languages, such as openCypher, were developed for querying and processing graph data. These languages usually offer a more intuitive way to express graph queries than SQL-like languages. However, most enterprises still store their data in traditional relational databases, which necessitates loading their data to graph databases. This is often impractical or infeasible for production databases. Our goal is to allow using expressive graph query languages and leverage the performance of existing relational databases while avoiding the overhead of transferring the data between different systems. To this end, we developed a transpiler which can transform openCypher graph queries into SQL.

Comparing the performance of database systems requires standard benchmarks. For relational databases, this is fulfilled by the benchmarks of the Transaction Processing Performance Council. Due to the relative immaturity of graph databases, there is only a limited number of benchmarks available for graph query workloads. We joined the development of the LDBC (Linked Data Benchmark Council) Social Network Benchmark. We reworked and significantly improved existing implementations of the benchmark, and also implemented the queries in the SPARQL language for semantic databases. We performed a thorough evaluation and detailed analysis of database systems using various data models (relational, graph, and semantic).