CS: 257: Project Proposal

Team Members

- Ashish Khanchandani
- Mit Jain
- Cajetan Rodrigues

Description

We aim to determine the difference between choosing a traditional RDBMS and a graph-based NoSQL database. We would like to execute a comprehensive comparison by using a dataset and querying the same data in both schemas across different categories. We would use Neo4j for the graph-based NoSQL database and MySQL for our traditional RDBMS comparison purposes. We would compare the performance of various operations like search, pattern matching, recursion and aggregation. Neo4j is touted to be one of the best graph bases systems in the industry; well known for its execution speed and the benefits that come with having a graph structure with nodes and edges to model the data effectively. MySQL is a very popular and widely used RDBMS. The purpose is to show which is better and how significant of a difference it makes if either database is chosen.

Motivation

Graph databases revolutionized the way data is stored and processed: By representing data as nodes and edges in a graph, they enable us to uncover insights that would be impossible to detect or require complex and expensive join operations with traditional relational databases. They allow us to efficiently navigate through vast and intricate networks of data, making them invaluable tools for applications ranging from e-commerce to scientific research. The research is motivated by the comparison of MySQL vs Graph databases and suggesting which database is suited under which scenarios.

Previous Work

Various studies have compared the performance of MySQL and Neo4j graph databases for different types of queries and datasets. Some studies have found that Neo4j performs better than MySQL in terms of query speed, while others have found that MySQL is faster and more memory-efficient. The types of queries tested include search/selection, aggregate, recursive, pattern matching, clustering, and path queries. The studies also explore the use of graph databases in various domains, such as social network analysis, web-based applications, IoT data management, and CRM systems. Overall, the studies suggest that the performance of graph databases is better than that of conventional databases for certain types of queries and datasets.

Methodology

Create 4 types of queries based on the 4 categories below

1. Selection
2. Recursion
3. Aggregation
4. Pattern Matching

Compare the performance of queries on MySQL and Neo4j databases for each of the 4 queries. Suggest scenarios and use cases where SQL and Graph-based databases can be best used.

Project Plan

Milestone 1 : (60% complete)

- Setup MySQL and Neo4j Databases
- Load datasets into both databases
- Apply Indexing
- Execute queries
- Evaluate the performance of at least 2 categories of queries

Milestone 2:

- Evaluation of the other 2 categories
- Further optimizations if possible via DB techniques
- If bandwidth exists, try performance tests with one more data set

Milestone 3:

- Prepare final report and source code ready for submission

Resources needed

1. MySQL Server (Local/Cloud) & Neo4j Aura Server (Local/Cloud)
2. Career Village Dataset
   https://www.kaggle.com/competitions/data-science-for-good-careervillage/overview

Citations/References:

1. Query-based Performance Comparison of Graph Database and Relational Database
2. A comparison of a graph database and a relational database: a data provenance perspective
3. Survey of graph database models
4. Comparative Analysis of Relational And Graph Databases
5. Comparative Analysis of Relational And Graph Databases
6. A Comparative Study of Relational and NonRelational Database Models in a Web- Based Application