# Introduction

One of the benefits of using NoSQL databases has usually been performance over traditional SQL databases. With an appropriate use case, NoSQL databases might offer significant performance benefits. Key-value based queries is one typical use case when NoSQL can give performance benefits. A graph database is one of the NoSQL database types. As the graph model consists of nodes and edges, it should be theoretically more optimal for relational queries. While in SQL database multiple tables have to be joined for a relational query, in graph databases relational information can be queried by navigating through the graph.

This study compares the relational query performance of SQL database MariaDB and graph database Neo4J. The databases are tested with various queries including a relation, aggregation, key-based and recursive query. There are already several studies where graph databases, especially Neo4J, have been compared to traditional SQL databases. In many cases the results often show better performance in favour of graphdb.

In this study we are going to compare Neo4J with the latest version of MariaDB and old version of MySQL. We were partially able to repeat similar results when comparing Neo4J with MySQL, showing performance benefits of Neo4J. However in many of our results MariaDB seemed to be radically better indicating how SQL database can offer good performance compared to the new graph based one.

# Related study

Various studies about graph database performance exist including [graph1], [graph2], [graph3] and [graph4]. [graph1] conducts the qualitative and performance comparison of 12 open source graph databases. Four algorithms (single source shortest path (SSSP), connected components (SV), PageRank (PR) and update) are computed on these databases and results for each database are shown with bar charts. The top five performers in these tests are STINGER, MTGL, Boost, Graph and NetworkX.

In [graph2] Neo4J query performance is tested with two datasets and two kinds of queries. The dataset consists author nodes and paper nodes. The relationship between these nodes is called ref. The first query is “find coAuthor of a paper written by an author” and the second query is “find a paper written by an author”. The queries are done three ways. With Java APi, with TraverserFrameWork and with Cypher. This study shows the clear differences between the three ways. Both Core Java API and TraverserFrameWork offering faster response time compared to Cypher.

[graph3] compares tuned Oracle 11g and Neo4J 3.03 Community Edition. Performance of the databases is evaluated with ten different count(\*) queries. Many of the queries also do some table joins. Physical database tuning techniques are used for Oracle and they seem to improve performance. However Neo4J performs still better in all of the tests.

In [graph4] Neo4J version 1.8 performance is tested with different backend solutions. Neo4J is benchmarked as embedded with native object access, as a dedicated server trough RESTful Web Services, with embedded Cypher queries, with Cypher trough REST optimized for remote execution and with Gremlin queries trough REST. MySQL version 5.5.27 is also included with Japa Persistence API based backend. Queries are done using Cypher, Gremlin and SQL query languages. The test data consist of people related data. Relational test queries are executed such as friends of friends. The results show embedded setups performing better than the REST based ones. This is due network overhead caused by the REST queries. Embedded Neo4J setups also perform better than JPA MySQL setups.

There also exist previous studies that compare Neo4J and MySQL [graph5] [graph6]. Both of these articles demonstrate the superiority of Neo4J over MySQL. Article [1] is from the year 2010 and compare MySQL Community Server version 5.1.42 and Neo4J version 1.0-b11. Article [2] is from the year 2012 and compares MySQL version 5.1.41 and Neo4J Community version 1.6. In the first article, the graph database ßis stored into a relational database as nodes and edges. The second article stores the database into a relational database in a relational form. In this study, the relational database also contains the database in relational form so the second article is more relevant to this study.

# Warehouse database

The test database is called the warehouse. The database has 10 tables. The basic tables are customer, invoice, target, work, worktype. These tables contain data about the articles used in the warehouse database. Relational information between the tables is stored in worktarget, workinvoice, useditem and workhours.

Kuva, joka sisältää kohteen näyttökuva, tie

Kuvaus luotu automaattisesti

Figure 1: Warehouse database in relational format



Figure 1: Warehouse database in relational format (Dia diagram)

At graph format edges are used to represent the relationships. For N:M relationships, bidirectional edges are used. Customer, invoice, target, work, worktype are represented as nodes. The edges between the nodes are PAYS between customer and invoice, CUSTOMER\_TARGET between customer and target, WORK\_TARGET between work and target, WORK\_INVOICE between work and invoice, WORKHOURS between work and worktype, USED\_ITEM between work and item.Kuva, joka sisältää kohteen pieni, varuste, kaulakoru

Kuvaus luotu automaattisesti

Figure 2: Warehouse database in graph format (from Neo4J)



Figure 2: Warehouse database in graph format (Dia diagram)

# Data generation

The test data is generated using a Java program. The program uses sample data that is based on openly available name and address data sets [3], [4]. The sample data is used when generating customer and target names. The generation process is divided into three parts. Items and work types should be generated first, then work and customer data. The Java program has threaded classes for each part. Multiple threads can be used to install data as well as factors for related data.

When generating work data, amount of the related worktypes and items can be defined by setting worktypefactor and itemfactor. When generating customer data, the amount of related invoices, targets and work can be defined by setting invoicefactor, targetfactor and workfactor. When generating work and customers, the given amount of work and customers are generated. For each work, the given factor of relations to worktypes and invoices are generated. For each customer given amount of invoices and targets are generated. The generator will also generate workinvoice and worktarget relationships based on the given factor.

# Test queries

The query tests contain different relational queries, an aggregation query and a recursive query. These queries test different capabilities of the databases. As both relational and graph databases are tested, the queries are in SQL and Cypher form. The test queries also aim to query something related to a practical use case. One of the most important usages of the database is to get the invoice price. The schema does not store invoice prices explicitly. The price has to be calculated based on the amount of the workhours and amount of items used. The short query and the long query are the subqueries for calculating this price. The aggregation query will calculate the whole price. A query with defined key will get the work for certain invoice id and recursive query will query all the interrelated invoices. One practical example is that customer has not paid the whole bill and there will be additional invoices based on the same invoice.

## Short query, workhours price

This query calculates the price of work hours. Different work types have different prices so the prices vary based on the hours for each work type. This query tests performance capabilities when querying relational data.

SELECT (price \* hours \* workhours.discount) as price FROM worktype,workhours,work WHERE worktype.id=workhours.worktypeId AND workhours.workId=work.id

MATCH (wt:worktype)-[h:WORKHOURS]->(w:work) RETURN (h.hours\*h.discount\*wt.price) as price

## Long query, work type price

This query is an extended version of the query that queries workhours prices. This query also includes item prices for each work which is summed to the workhours prices. As items are also included, the longer relational query is needed.

SELECT (price \* hours \* workhours.discount) + (purchaseprice \* amount \* useditem.discount) as price FROM worktype,workhours,work,useditem,item WHERE worktype.id=workhours.worktypeId AND workhours.workId=work.id AND work.id=useditem.workId AND useditem.itemId=item.id

MATCH (wt:worktype)-[h:WORKHOURS]->(w:work)-[u:USED\_ITEM]->(i:item) RETURN (h.hours\*h.discount\*wt.price)+(u.amount\*u.discount\*i.purchaseprice)

## Query with defined key, work of invoice

This query tests the performance of a relational query for a specific invoice. When an invoice id is known, querying relationships should be straightforward in a graph database. In a relational database, relation information should be queried from a relation table.

SELECT \* FROM invoice,workinvoice,work WHERE invoice.id=workinvoice.workId AND workinvoice.workId=work.id AND invoice.id=0

MATCH (i:invoice { invoiceId:0 })-[wi:WORK\_INVOICE]->(w:work) RETURN \*

## Aggregate query, invoice price

This query calculates the sum of a work price for each invoice. The query contains two subqueries. The results of these queries are joined and the sums of prices are aggregated based on invoice id.

SELECT q1.invoiceId AS invoiceId, sum(q2.price) AS invoicePrice FROM (SELECT invoice.id AS invoiceId, work.id AS workId FROM invoice, workinvoice, work WHERE invoice.id=workinvoice.invoiceId and workinvoice.workId=work.id ) AS q1, (SELECT work.id AS workId, SUM((worktype.price \* workhours.hours \* workhours.discount) + (item.purchaseprice \* useditem.amount \* useditem.discount)) AS price FROM worktype,workhours,work,item,useditem WHERE worktype.id=workhours.worktypeid AND workhours.workid=work.id AND work.id=useditem.workid AND useditem.itemid=item.id GROUP BY work.id) AS q2 WHERE q1.workId = q2.workId GROUP BY q1.invoiceId

MATCH (inv:invoice)-[:WORK\_INVOICE]->(w:work)<-[h:WORKHOURS]-(wt:worktype) WITH inv, w, SUM(wt.price\*h.hours\*h.discount) as workTimePrice OPTIONAL MATCH (w)-[u:USED\_ITEM]->(i:item) WITH inv, workTimePrice + SUM(u.amount\*u.discount\*i.purchaseprice) as workItemPrice RETURN inv, sum(workItemPrice) as invoicePrice

## Recursive query, invoices related to invoice id 100000

This is a recursive query that gets all the invoices related to a given invoice. As MySQL does not support Common Table Expressions they are not used.

SELECT id,customerid,state,duedate,previousinvoice FROM (SELECT \* FROM invoice ORDER BY previousinvoice, id) invoices\_sorted, (SELECT @pv := '100000') initialisation WHERE find\_in\_set(previousinvoice, @pv) AND length(@pv := concat(@pv, ',', id))

MATCH (i:invoice { invoiceId:100000 })-[p:PREVIOUS\_INVOICE \*0..]->(j:invoice) RETURN \*

# Query tests

## Test settings

The query tests were performed with MacBook Pro Laptop that has following specifications:

* macOS Catalina version 10.15.3
* 1,4 GHz quad core Intel Core i5
* 8 GB 2133 MHz LPDDR3
* Intel Iris Plus Graphics 645, 1536 MB

MySQL version 5.1.41, MariaDB version 10.4.11-MariaDB and Neo4J version 3.5.14 were installed on this computer. MariaDB driver version 2.1.2 and Neo4J driver version 4.0 were used. A dataset that contains 10000 customers, 100000 invoices, 100000 targets, 10000 works, 100000 items, 100000 worktypes, 100000 useditems, 100000 workhours, 1000000 workinvoices and 1000000 worktargets was generated.

## Test results

Each query test was executed with ten iterations. A list of results was formed after each query. The biggest and the smallest number was removed from the list and an average was calculated from the remaining numbers.

* + 1. Short query

Results for a short query can be found in table 1. In these results, Neo4J performs better than MySQL 5.1.41. However, MariaDB outperforms both.



Table : Results for short query

* + 1. Long query

Results for a long query can be found in table 2. Results show similar order to databases as in short query MariaDB outperforming both of the databases.



Table : Results for long query

* + 1. Query with defined key, work of invoice

Results for a query with defined key can be found in table 3. Neo4J would be expected to have the best performance with this query as when the key is known, relations can be found by navigating the path. However bot MySQL and MariaDB seem to perform better.



Table 3: Results for query with defined key

* + 1. Aggregate query, invoice price

An aggregation query is very heavy as it contains two subqueries. The performance tests took long to complete. In these tests, Neo4J outperformed MySQL and MariaDB outperformed both.



Table : Results for aggregate query

* + 1. Recursive query, invoices related to invoice id 100000

The recursive query lists all invoices related to invoice 100000. There were total 100 invoices to query. In this test, the order of databases is the same as in the first query tests Neo4J outperforming MySQL and MariaDB outperforming both.



Table : Results for recursive query

# Conclusions

The performance tests represented in this article clearly show the superior performance of relational database MariaDB. The results are consistent with previous studies showing Neo4J outperforming an old version of MySQL. However with the new MariaDB database the situation has changed in favour of a relational database gaining back the position when it comes to performance. While the graph model should be in theory more effective in relational queries, the new relational database seem to be still a viable option when it comes to performance.

[3] https://data.world

[4] <http://results.openaddresses.io/>

[graph1], McColl, Robert Campbell, et al. "A performance evaluation of open source graph databases." *Proceedings of the first workshop on Parallel programming for analytics applications*. 2014.

[graph2], Huang, Hongcheng, and Ziyu Dong. "Research on architecture and query performance based on distributed graph database neo4j." *2013 3rd International Conference on Consumer Electronics, Communications and Networks*. IEEE, 2013.

[graph3], Khan, Wisal, et al. "SQL Database with physical database tuning technique and NoSQL graph database comparisons." *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, 2019.

[graph4] Holzschuher, Florian, and René Peinl. "Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j." *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. 2013.

[graph5] Vicknair, Chad, et al. "A comparison of a graph database and a relational database: a data provenance perspective." *Proceedings of the 48th annual Southeast regional conference*. 2010.

[graph6] Batra, Shalini, and Charu Tyagi. "Comparative analysis of relational and graph databases." *International Journal of Soft Computing and Engineering (IJSCE)* 2.2 (2012): 509-512.