# Warehouse database

The test database is called 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

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 3: 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 [1], [2]. 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 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.

## 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 10 iterations. A list of results was formed after each query. Biggest and smallest number was removed from the list and an average was calculated from the remaining numbers.

* + 1. Short query



* + 1. Long query



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



* + 1. Aggregate query, invoice price



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



[1] https://data.world

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