**Bachelor Thesis**: Benchmarking Graph Databases – Conducting a Benchmark on Graph Databases with Industry 4.0 Data and Workloads

**Start**: 01.07.2017\* **End**: 01.11.2017\*

\*planned

**ABSTRACT**

Graph databases are a good way to store data from the shopfloor in the industry. To help the industry decide which database suits the best many have to be compared. To compare them workloads will be analysed as well as data structures. Then existing benchmarking software will be evaluated and extended to measure the important aspects for the industry. Lastly the benchmarks will be conducted with different graph databases and the results presented in this thesis.

**Research Question**

Are current graph databases able to deal with the arising Industry 4.0 data?

* Are they able to write the amount of data?
* What is the maximal load they can handle?
* Are they responsive unter load?
* Are they able to run analysis algorithms on the data?
* Are they able to perform reasoning on the data?

**AREA and PROBLEM**

Area: For companies, switching to Industry 4.0 means, digitalising and virtualising their workflows and production processes. If done correctly, a lot of data will become available for optimizing processes and machines, data that should not be thrown away. In order to use this data, it needs to be stored for future analysis or observation.

Since Industry 4.0 uses the Internet of Things paradigm every machine, worker, component and product has its own digital representation in the virtual copy of the factory. The relationship between these entities can be mapped to nodes and edges in a graph.

Therefore, it is natural to choose a graph to represent the virtual factory with digital twins as nodes and connections as edges.

Problem: With the increasing variety of graph databases available and each performing better at different tasks it is hard to find the best fitting one for a set criteria. For better comparison between many databases a benchmark could hand an advice which database would perform the best in a given scenario. Existing benchmarks do not implement a workload for writing data continuously into the database, which is a use case in Industry 4.0. Also, the data structure in this context is different from a social media graph used in most benchmarks.

Solution: The student will provide a benchmarking software for testing the throughput of write operations, responsiveness under different loads and the maximum load a database can handle.

**BACKGROUND of the THESIS**

Industry 4.0 is all about data and using it beneficially. By analysing the data, weak spots in the production can be found and fixed to save time and money.

To store this data the company needs an adequately performing database in both the aspects of writing their data from the manufacturing process and analysing it.

Because of the Internet of Things paradigm and a digital twin for every entity in the factory it is natural to us a graph database. Every machine and component is represented by a node and connections between two nodes represent a connection in the real world, for example a component is “produced\_in” a machine.

To choose a suitable graph database, benchmarks can offer a good advice about specific performance aspects. Existing benchmarks do not test continuous writing to the database. Also, the used data structure differs from the one in Industry 4.0.

**BASIC IDEA of the THESIS**

The idea of this thesis is to deliver a benchmarking software which delivers relevant results for the industry. The benchmark should test the databases with workloads derived from industrial workflows.

A realistic data structure should be used for those workloads.

This benchmark should aid the industry to decide which database suits their needs the best.

**DESCRIPTION of overall PROCEDURE**

The student will examine existing benchmarking software to decide which could be extended to meet the specified solution.

He has to analyse the industrial data structure and derive the key parameters for generating test data or represent this structure with his own model.

The workflows have to be acquired and the benchmark extended accordingly.

**SCIENTIFIC CONTRIBUTION**

This thesis will contribute the knowledge of suitability of graph databases for usage in the industry.

**TASKS**

The thesis groups into five parts.

Literature research, to get to know the benchmarks, databases to be tested and industry needs. This should take about four weeks.

Designing the workloads and data structure for implementation and evaluation, that should take about one week.

Evaluating existing benchmarking software for suitability of extension with data structure and workloads, this should take around two weeks.

Implementing the workload, data structure and setting up the system for evaluating should take about four weeks.

The conduction of the benchmark and evaluation of the results should take about one week.

Finally writing the thesis should take around four weeks.

Literature: 4 Weeks (25%)

Design: 1 Week (6,25%)

Examination: 2 Weeks (12,5%)

Implementation: 4 Weeks (25%)

Evaluation: 1 Week (6,25%)

Writing: 4 Weeks (25%)

Total: 16 Weeks

**CORE TASKS**

The core tasks of the thesis include the following topics.

Evaluate existing benchmarking software

* Has parameterized data generator
* Can export data for reproducible testing
* Can write via a workload

Designing and evaluating a data structure for the test

* Designing of a structure
* Analysing this structure (key attributes of the graph)
* Figuring out how to create an appropriate graph

Designing the workloads

* Continuously writing
* Responsiveness under different loads
* Maximum load before failure

Set up testing environment

* Only single node or distributed or test both
* Dependencies on environment

**RISKS**

Risk Description: The data generator can’t export the produced dataset.

Risk Impact: An exporting and importing mechanism has to be implemented

Risk Likelyhood: Likely

Preventive Measure: None

Reactive Measure: Implement a program to export and import the data for testing.

Risk Status: -

Risk Description: The examined data structure can’t be represented with the current available generators.

Risk Impact: A data structure has to be implemented which costs time.

Risk Likelyhood: Moderate

Preventive Measure: None

Reactive Measure: Implement a reduced data structure to represent the most imported aspects.

Risk Status: -

Risk Description: The workloads from the underlying benchmark can’t be used for updating the database.

Risk Impact: Benchmark could not be done that easily.

Risk Likelyhood: Moderate

Preventive Measure: None

Reactive Measure: Implement another structure for realising the updating workload.

Risk Status: -