Review report of a final thesis

Czech Technical University in Prague **Faculty of Information Technology**

Student:	Ing. Jaroslav Ramba
Reviewer:	Michal Bachman, MSc.
Thesis title:	Indexování struktur v grafovém DB stroji neo4j II
Branch of the study:	Web and Software Engineering (Master, in Czech and in English)

Date: 14. 6. 2015

Evaluation criterion:	*The evaluation scale: 1 to 5.*
1. Difficulty and other comments on the assignment	***1 = extremely challenging assignment,*** *2 = rather difficult assignment,* *3 = assignment of average difficulty,* *4 = easier, but still sufficient assignment,* *5 = insufficient assignment*

Criteria description:
Characterize this final thesis in detail and its relationships to previous or current projects. Comment what is difficult about this thesis (in case of a more difficult thesis, you may overlook some shortcomings that you would not in case of an easy assignment, and on the contrary, with an easy assignment those shortcomings should be evaluated more strictly.)

Comments:

The assignment has been extremely challenging for three reasons. First, the assignment was quite open-ended - indexing in databases is a broad topic and it is broader still in graph databases, which have an arguably richer data model. Secondly, to the best of my knowledge, indexing of patterns in graph databases and Neo4j in particular has not been addressed in any academic research to date. Finally, Neo4j is a relatively new technology; hence, the practical part of the assignment required a deep understanding of how to integrate with the technology, which is often documented only in its (open-sourced) code.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
2. Fulfilment of the assignment	***1 = assignment fulfilled,*** *2 = assignment fulfilled with minor objections,* *3 = assignment fulfilled with major objections,* *4 = assignment not fulfilled*

Criteria description:
Assess whether the thesis meets the assignment statement. In Comments indicate parts of the assignment that have not been fulfilled, completely or partially, or extensions of the thesis beyond the original assignment. If the assignment was not completely fulfilled, try to assess the importance, impact, and possibly also the reason of the insufficiencies.

Comments:

The assignment was fulfilled as specified. Shortcomings of the solution are clearly documented. Specifically, the absence of a reliable protocol that keeps the index consistent with the database in case of failures was addressed by pointing out the lack of two-phase commit.

Evaluation criterion:	*The evaluation scale: 1 to 4.*
3. Size of the main written part	*1 = meets the criteria,* ***2 = meets the criteria with minor objections,*** *3 = meets the criteria with major objections,* *4 = does not meet the criteria*

Criteria description:
Evaluate the adequacy of the extent of the final thesis, considering its content and the size of the written part, i.e. that all parts of the thesis are rich on information and the text does not contain unnecessary parts.

Comments:

The size of the written part meets the criteria. The text contains some unnecessary parts. For example, evaluation of other graph databases could have been omitted, since it adds no value to the reader (assignment was specified for Neo4j). Likewise, mentioning other hash table implementations adds little value, since it is not properly explained, why MapDb has been chosen.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
4. Factual and logical level of the thesis	*80 (B)*

Criteria description:
Assess whether the thesis is correct as to the facts or if there are factual errors and inaccuracies. Evaluate further the logical structure of the thesis, links among the chapters, and the comprehensibility of the text for a reader.

Comments:

The thesis is logically well-structured and comprehensible for the reader.

Especially in the first part of the thesis, however, there are a few factual inaccuracies. The author could have spared himself explaining how data is stored on disk in Neo4j in the first place by stating that data is stored in doubly-linked lists. The details presented contain an inaccuracy in saying that 1 byte (instead of bit) is used for a used/not used flag (2.2.3.2), and uses an out of date detailed data layout on disk (figure 2.3 is for Neo4j 1.9, but Neo4j 2.2 was used for the thesis).

Other comments:

- Section 3.1 is particularly good and readable
- Section 3.2 contains no explanation at all why a particular data structure (hash map) was chosen and what is the reason for the design of its key contents. No alternatives are suggested
- The fact that there are relationship types, directions, node labels, and node and relationship properties in Neo4j is barely mentioned in the design of the index, yet these building blocks are used by virtually every Neo4j user
- Cypher query in section 4.4.5 could use some explanation
- the lack of two-phase commit is purely stated, but there is no discussion as to why it may be needed in the first place and what are the implications of it lacking

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
5. Formal level of the thesis	*90 (A)*

Criteria description:
Assess the correctness of formalisms used in the thesis, the typographical and linguistic aspect s, see Dean's Directive No. 12/2014, Article 3.

Comments:

Apart from a few colloquial expressions, the formal level of the thesis is good.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. Bibliography	*85 (B)*

Criteria description:
Evaluate the student's activity in acquisition and use of studying materials in his thesis. Characterize the choice of the sources. Discuss whether the student used all relevant sources, or whether he tried to solve problems that were already solved. Verify that all elements taken from other sources are properly differentiated from his own results and contributions. Comment if there was a possible violation of the citation ethics and if the bibliographical references are complete and in compliance with citation standards.

Comments:

The author used a good amount of relevant sources, especially considering the fact that literature on the topic is sparse. For parts with no available literature available, he used relevant online content, presentations, and discussion forums, and even travelled to London to meet the creators of Neo4j in person and ask questions and opinions.

There are some minor shortcomings in referencing, for example, I would not consider reference [21] to be a reliable source of information. No references mentioned in the first paragraph of chapter 3.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
7. Evaluation of results, publication outputs and awards	*90 (A)*

Criteria description:
Comment on the achieved level of major results of the thesis and indicate whether the main results of the thesis extend published state-of-the-art results and/or bring completely new findings. Assess the quality and functionality of hardware or software solutions. Alternatively, evaluate whether the software or source code that was not created by the student himself was used in accordance with the license terms and copyright. Comment on possible publication output or awards related to the thesis.

Comments:

Good results were achieved in the sense that the solution proved to be feasible, bring value by decreasing query times for certain queries by orders of magnitude, whilst having minimal footprint and impact on transactional processing.

The thesis is would be a good starting point for further research on the topics. Before publishing, different shapes should be tested and, more importantly, implications of labelled property graph model (i.e., the fact that nodes have labels and properties and relationships have types and properties) should be considered.

The software was created by the student and is of good quality for academic research purposes. For production use, it would have to be undergo more rigorous testing process.

A section entitled "Further Research" would be useful, indicating in which directions researchers building on this work might want to take.

Evaluation criterion:	*No evaluation scale.*
8. Applicability of the results	

Criteria description:
Indicate the potential of using the results of the thesis in practice.

Comments:

The results, especially the design of the index and the measurements, are practically applicable for people looking to build a pattern index in graph databases. It is not unlikely that the ideas in this thesis, in one way or another, will eventually make their way into the core of graph database implementations.

Evaluation criterion:	*No evaluation scale.*
9. Questions for the defence	

Criteria description:
Formulate any question(s) that the student should answer to the committee during the defence (use a bullet list).

Questions:
- Why was hash map used for the index implementation, if only the key (not value) is used? Could you have used a set?
- What implication would the fact that there are relationship types, directions, node labels, and node and relationship properties in Neo4j have on your research?
- Could you explain the query in section 4.4.5?
- You talk about the lack of two-phase commit. Why may it be needed in the first place? What are the implications of it lacking?

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
10. The overall evaluation	*85 (B)*

Criteria description:
Summarize the parts of the thesis that had major impact on your evaluation. The overall evaluation **does not** have to be the arithmetic mean or any other formula with the values from the previous evaluation criteria 1 to 9.

Comments:
The author has managed to scope the extent of this open-ended research very well. Within that scope, he worked independently with great dedication and a steady progress in "design-prototype-learn" cycles, resulting in a valuable and practically applicable piece of research. The written part of the thesis doesn't do justice to the research performed and the work carried out, and could be improved upon, as described above.

Signature of the reviewer: