Review report of a final thesis

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

Student: Bc. David Příhoda
Reviewer: Ing. Jan Motl
Thesis title: Distributed Conversion of RDF Data to the Relational Model
Branch of the study: Knowledge Engineering

Date: 8. 6. 2017

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 is not easy as it requires taming of three complex systems: Spark, relational databases and triplestore databases.
Furthermore, three aspects of the solution were addressed at once: completeness, usability and performance.

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:

Satisfied

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:

Since this is an implementation thesis, the length of the text is appropriate. The text is complemented with many useful figures.

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

4. Factual and logical level of the thesis

95 (A)

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:

I just do not understand why terms "Country" and "Eligibility" are duplicated in the "junction tables" in figure 6.2. But that can be a quirk of Tableau software.

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

5. Formal level of the thesis

88 (B)

Criteria description:
Assess the correctness of formalisms used in the thesis, the typographical and linguistic aspect s, see Dean's Directive No. 14/2015, Article 3.

Comments:
IRI abbreviation is not explained in the text, although used frequently.

The term "relation" is used in three distinct meanings in the thesis:
1) Table
2) Relationship (as a type of tables in Entity-Relationship Diagrams)
3) Reference (as in a foreign key constraint)

While all these usages are correct (at least by Codd's and Chen's terminologies), when present together in a single document, it decreases the clarity of the text.

Similarly, term "relationship" is used in two meanings:
1) As a type of a table in the Entity-Relationship Diagram
2) As a link

While the assignment touches multiple domains, making it significantly more difficult to assign unique terms to distinct concepts, a reader should be at least warned about the presence of the overloaded terms.

Small details:
1) Drugbank -> DrugBank
2) RDF2RDB[57] -> RDF2RDB [57]
3) can be seen on figure -> can be seen in figure

Since these are all my complaints and I enjoyed the illustrations -> B+.

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

6. Bibliography

90 (A)

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:
In references:
[4]: pp. 1-1 -> pp. 38-43
[13] I am not sure the document has page 2004

I do not have any more complaints -> A.

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

7. Evaluation of results, publication outputs and awards

80 (B)

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:
It is argued in section 1.2.2 that relational stores have more predictable query runtime than RDF stores. This is not backed by measurements or literature reference.

I find it regrettable that the matches between:
1) the original relational schemata
2) the generated relational schemata from RDF2X
were not evaluated.

Nevertheless, the speed of the conversion and sparsity of the results are well evaluated. And a nice breakdown of the runtime is provided in figure 5.2 -> B.

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:
Conversion of RDF data to relational data is not common as there are simply not too many RDF sources. However, if you need to convert RDF data to relational data, you generally need a tailored solution that can scale. The developed tool provides the solution and is awaiting approval for open-source release by MSD IT Prague.

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:
1) How do the runtimes of the SQL queries compare to the runtimes of SPARQL queries on tasks listed in chapters 6 and7?
2) Did the conversion of RDF data to relational format fulfill the expectations?

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
10. The overall evaluation	*87 (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:
Very good delivery for an above averagely difficult topic.

Signature of the reviewer: