FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Supervisor's statement of a final thesis

Student:	Bc. Marek Tornóci
Supervisor:	Ing. Jan Trávníček, Ph.D.
Thesis title:	Metadata extraction, parsing, and dataflow detection in Snowflake sql dialect
Branch of the study:	Web and Software Engineering

Date: 25. 8. 2020

Evaluation criterion:	*The evaluation scale: 1 to 4.*
1. 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 submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

Comments:

The thesis assignment required the student to investigate a possible approach to extract metadata from the Snowflake database and from the scripts in order to create a representation of dataflow in those scripts. Both the thesis and the attached code fulfil the assignment.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
2. Main written part	*90 (A)*

Criteria description:
Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies? Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 26/2017, Art. 3. Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

Comments:
The thesis is of appropriate length, with all chapters relevant, and is written in English. The thesis may serve as a basis of the attached code documentation.

Factual issues:
- It would make sense to relate AST examples to grammar fragments in ANTLR format to see details behind differences in figures 1.3 and 1.4. Also, nodes are said to be added but, for instance, the "FROM" node disappears in the transition from figure 1.3 to figure 1.4.
- In general, I miss the most, at least, a semiformal definition of grammars as they are the fundamental tools in any formal language processing.
- The "possible ways of metadata extraction" section is only vague about the DDL extraction approach and it is not clear how to extract DDLs from Snowflake database.
- Despite it being a pedantic issue, I have to mention that the example in the Source Code 2.4 could be more meaningful with the alias_t1 view used in the inner select.
- Functional requirement section 2.4.1 states that all required DDL statements are to be extracted and stored, however, I was unable to find which DDL statements are required.
- It feels that the statement "Because only a valid SQL script can be used as an input into the parser, ..." actually builts on a functional requirement "Only valid scripts will be parsed", which is not in the list of the functional requirements.

Questions regarding the text:
- Would the DDLs of functions help with the function parameter issue introduced in Section 2.2.2.5, or does the function's DDL exhibit the same issue?
- Regarding Source Code 2.6, can the aliasing be achieved using subselect and aliasing it?
- Is there any way to specify the internal structure of JSON stored within a column of variant type or everything is dynamic?

Typography and wording issues:
- Description of the first chapter in the "Goal of the thesis" does not seem to have a correct grammatical structure. (Additionally, the second chapter does not analyze, it is rather a documentation of analysis.)
- Sometimes the sentences are too long, which makes them harder to read. (For instance, the second paragraph of section 1.1 is a single sentence.)
- Introduction of table database objects refers to stages that are introduced later.
- The title of Source Code 2.4 and 2.12 is split from the listing by a page break.
- Incorrectly placed comma in "In the example, 2.9 we can see".
- Bullet lists that are part of sentences do not have the expected grammar structure.
- Text overflows to the right on pages 36, 44, 46, 47, and 48.
- Figures 6.1 and 6.3 are bitmap even though they are generated by Graphviz which is able to produce vector images.
- Incorrect grammar structure "The class contains implemented a following set ...".

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
3. Non-written part, attachments	*99 (A)*

Criteria description:
Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

Comments:
The code follows the Manta's internal guidelines, it is well commented and appropriately tested.

I have some minor comments regarding the ANTLR sources:
- Sometimes, a += operator is used where = would suffice.
- It seems that the second alternative in with_select_statement_no_master is unnecessary.

What is the interpretation of "1e" code fragment, is it supposed to be understood by the Snowflake SQL dialect as integer and identifier (as implemented) or as float (some SQL dialects actually do that)?

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

Criteria description:
Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

Comments:
The provided source codes are already incorporated into the Manta codebase and will gradually be improved to process all statements of the Snowflake SQL dialect.

Evaluation criterion:	*The evaluation scale: 1 to 5.*

5. Activity and self-reliance of the student	*5a:* ***1 = excellent activity,*** *2 = very good activity,* *3 = average activity,* *4 = weaker, but still sufficient activity,* *5 = insufficient activity* *5b:* ***1 = excellent self-reliance,*** *2 = very good self-reliance,* *3 = average self-reliance,* *4 = weaker, but still sufficient self-reliance,* *5 = insufficient self-reliance.*

Criteria description:
From your experience with the course of the work on the thesis and its outcome, review the student's activity while working on the thesis, his/her punctuality when meeting the deadlines and whether he/she consulted you as he/she went along and also, whether he/she was well prepared for these consultations (5a). Assess the student's ability to develop independent creative work (5b).

Comments:
The cooperation with the student was excellent.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. The overall evaluation	*97 (A)*

Criteria description:
Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.

Comments:
The issues in the text and code are of minor importance, therefore I recommend the thesis for defence and recommend to grade it with 97 points, i.e. grade A (excellent).

Signature of the supervisor: