FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Supervisor's statement of a final thesis

Supervisor:	Ing. Jan Trávníček, Ph.D.
Student:	Bc. Lucie Procházková
Thesis title:	Data flow analysis of scripts in Databricks SQL dialect
Branch / specialization:	Computer Science
Created on:	4 June 2023

Evaluation criteria

1. Fulfillment of the assignment

▸ [1] **assignment fulfilled**
[2] assignment fulfilled with minor objections
[3] assignment fulfilled with major objections
[4] assignment not fulfilled

The thesis aims to provide a new scanner and dataflow generator for the SQL scripts used within the Databricks system that is able to process a substantial subset of statements relevant to the subsequent analysis. The solution is supposed to cover parsing (lexical and syntactic analysis) and dataflow generation (semantic analysis based on static information). Last but not least, the solution is to be covered with unit tests.

I declare the solution fully fulfills the assignment.

2. Main written part

92 / 100 (A)

The thesis text is well-structured and comprehensively formulated. The chapters follow a logical ordering that is suitable and are information rich. I have identified some issues with the terminology used regarding theoretical background formalisms or notions, which, however, don't decrease the readability or understandability of the text. Grammar generates language; the initial symbol of regular grammar can be on the right-hand side of other rules, given that it cannot be rewritten to epsilon, to name a few.
I would appreciate a description of the node types in the dataflow graph introduction.

Code snippet 18 understandably shows an incomplete grammar of a select statement, which makes it a bit unclear: What is a clause? Is the VALUES clause the same as the WHERE clause? Where are the typical resultset combining operations like union, intersect, etc.?

There are a few typography issues, FIRST - - FIRST and FIRST - - FOLLOW; code snippet 18 seems to have inconsistent formatting, indentation, highlighting, and syntax. Page 62 contains a weird line break in the 5th paragraph.

All in all, the thesis text provides a solid, self-contained description of what was done while working on implementing the required scanner and dataflow generator (in the Manta terminology), and what were the design choices made.

3. Non-written part, attachments 95 /100 (A)

The implementation was directly contributing to the source code repository of Manta. Individual commits were reviewed before merging into the main development branch. The code is, as of this time, almost ready to be used as a minimal viable product implementation and thus given to the customers.

There, however, still are some gray areas to implement and improve. For instance, there are statements that do not impact the generated lineage, which the scanner cannot parse or parse completely. Also, it was recently discovered that the actual syntax of the undocumented select statement is more complicated. A refactoring of the code is ongoing, and chances are the redesign is going to land in the main development branch prior to the time of the defense.

4. Evaluation of results, publication outputs and awards 100 /100 (A)

The demand for cloud-based data storage and processing solution is increasing, and with it, the need for knowledge about the data lineage within such a system follows. The lineage is similar to one already available for many on-premise installed database systems, however, the Databticks cloud-based solution allows easier interconnection of other technologies, programming languages, etc. within the system.

5. Activity of the student

- ▸ [1] **excellent activity**
  [2] very good activity
  [3] average activity
  [4] weaker, but still sufficient activity
  [5] insufficient activity

The development of the solution, which lasted over many months (unsurprisingly, given the scope of the task), was tackled with equivalent activity.

6. Self-reliance of the student

- ▸ [1] **excellent self-reliance**
  [2] very good self-reliance
  [3] average self-reliance
  [4] weaker, but still sufficient self-reliance
  [5] insufficient self-reliance

There have been many consultations, but they were always productive and topic-focused.

The overall evaluation

95 /100 (A)

To summarize the report, I would like to repeat that the thesis is solid both from the point of view of the text and supplementary implementation. There is a small number of issues that are not anyhow crucially decreasing the quality of the thesis.

Instructions

Fulfillment of the assignment

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.

Main written part

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. 52/2021, 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.

Non-written part, attachments

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.

Evaluation of results, publication outputs and awards

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.

Activity of the student

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.

Self-reliance of the student

From your experience with the course of the work on the thesis and its outcome, assess the student's ability to develop independent creative work.

The overall evaluation

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.