FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Review report of a final thesis

Student:	Ing. Vitalij Kozlov
Reviewer:	Ing. Tomáš Vondra, Ph.D.
Thesis title:	Real-time Data Stream Processing System
Branch of the study:	Web and Software Engineering

Date: 16. 1. 2021

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 assignment was fulfilled completely.

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

2. Main written part

75 (C)

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 text of the thesis is well structured. A minor factual reservation about the theoretical part: The other possibilities of running Spark before Kubernetes are also container orchestrators. Therefore, the inclusion of the chapter about advantages of containers over VMs doesn't make sense. Also: There are several references to the CAP theorem, which is not explained nor as an abbreviation nor as a concept.
The English language is excellent as far as I can judge as non-native speaker.
Formally, I have a big reservation to citation style. The text mixes Harvard and IEEE citations styles. In some passages, citations are in square braces [10], in other as Author 2018 and in some, there are both in the same sentence. This is very unusual.
The cited resources themselves are of high quality with a great number of books on the topic.

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

3. Non-written part, attachments

90 (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:
I didn't have access to the appendices of the thesis, but I don't think I need them to judge that the implemented solution is workable as a proof of concept and I have no reason to think that it is not original.

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

4. Evaluation of results, publication outputs and awards

50 (E)

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:

My greatest resevation is to the applicability of the results. The thesis satisfies the assignment but the assignment itself is something akin to Hello World of stream processing. I have seen demos such as this countless times. The chosen approach with the Lambda architecture is correct, but the design pattern is from 2013 and I have already seen a few production systems use it, so there is no novelty in it. The thesis works quite linearly towards the solution, mainly in the technical part. I miss more discussion about other technologies, e.g. Why Spark Streaming and not Storm? Why exactly Cassandra? Or, alternatively, I would welcome a Related works part which would show similar solutions of others and compare them. As it is, is has most value to its author, who undoubtedly learned a lot while studying the topic.

Evaluation criterion: *No evaluation scale.*

5. Questions for the defence

Criteria description:
Formulate questions that the student should answer during the Presentation and defence of the FT in front of the SFE Committee (use a bullet list).

Questions:

The possibility of running Spark on Kubernetes is interesting in that you can also run other components on the same cluster, which was more limited with YARN. However, in production deployments of Big Data systems, different components are often deployed on hardware optimized for them. How would you, e.g., ensure that Kafka gets a dedicated storage for its data directory with fast linear writes?

EFS is designed as a shared consistent filesystem. Why didn't you choose S3 for long term storage of immutable objects? Wouldn't it be cheaper and more scalable? The data must not necessarily be on the same storage as checkpoints. Are there other options for storing Spark checkpoints than a file system?

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

6. The overall evaluation

75 (C)

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 theoretical part is a good summary of the best practices in stream processing. I would have liked more width in the technical part and more discussion of the different options. The implementation is a good proof of concept.

Signature of the reviewer: