FAKULTA INFORMATIKY
Masarykova univerzita

# POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

| | |
|---|---|
| Název práce: | *Leverage Data Quality Improvement for Big Data Analytics* |
| Autor práce: | *Ahmed Mansour Nasralla Taha Abdalla, učo 464106* |
| Vedoucí práce: | *Mouzhi Ge, Ph.D.* |
| Oponent: | *Bruno Rossi, Ph.D.* |

---

The thesis is focused on providing a Data Quality Improvement Framework (DQIF) that can enhance big data analytics. After discussing data quality problems from literature, the DQIF is proposed and evaluated in an experimental setting with real data.

The thesis is composed of five chapters and one appendix:

- Introduction: defines well the context, the problem statement and the research questions / goals. I like the idea of the research design as a figure, maybe this thesis is not so complex to warrant such use, but it is surely a worthy addition.
- Literature Review: very extensive, discusses data quality attributes and several existing frameworks. Very useful are the tables that compare the different aspects covered by literature. Only comment, maybe too much from Eppler's article is used in some sections (in the sense that they are entirely based on one reference), but that's a minor comment.
- Data Quality Improvement Framework: defines nicely the overall idea of a data quality framework and defines its characteristics.
- Experiment, Results and Evaluation: follows the previous chapter and gives a good idea of the application of the framework to a dataset from Transferwise.
- Appendix: contains the source code of the implemented scripts for the framework. Nice idea to add some in-text comments to describe the fragments of code. However, the code could have been attached also as supporting material for the thesis.

The thesis is excellent in terms of elaboration and level of details. Some issues I have noticed:

- Maybe I have missed it, but is the Transferwise dataset available online, or where was the data gathered from? Has an external researcher the possibility to replicate the overall approach in the experimental chapter?
- Citation style is a bit strange in some places: better not to start with *"[39] says that ..."*, rather *"x et al. [39] writes that.."*
- There are a couple of paragraphs in which text is copied from the reference (not correctly cited / quoted), e.g. "*The Accuracy Dimension is about assessing the quality of corporate data and improving its accuracy using the data profiling method....*". Given the amount of text of the thesis, this was probably just the student forgetting to rewrite those parts.
- There are also some mistakes throughout the thesis, e.g., *"Throw this section"* (p.19) - maybe *through*?
- There could be also aspects to discuss in the analysis. I want just to focus on one: Fig. 25 is reporting the accuracy over the k(10) folds. Are you sure there are no issues with NB? Seems to have a huge variability (high std dev). If not a mistake, can you give an explanation?

My curiosity is about the word "Big" in the thesis title. Big Data is mentioned, but is not extensively discussed, rather all is based on data analytics. So one question for the student: what would happen if you drop the word "Big" from the title? Would the thesis content change? Or in other terms: What are the differences (if any) for the application of DQIF to "data analytics" or "big data analytics"?

Considering the aforementioned comments, I would recommend to accept the work as Master Thesis of the Faculty of Informatics, with the proposed assessment of ***"Excellent"* (A)**.

Brno June 3, 2018 Bruno Rossi, Ph.D.