MUNI
FACULTY
OF INFORMATICS

MASTER THESIS
READER'S REVIEW

Title: *Deploying Data Lake for Big Data Management*

Author: *Tomáš Chomo, učo 475924*

Supervisor: *doc. Mouzhi Ge, Ph.D.*

Reader: *Bruno Rossi, Ph.D.*

The goal of the thesis is to propose a decision model for the adoption of data lakes within a big data context. The thesis is composed of five chapters (six counting the introduction chapter). The goals set in the thesis assignment seem to have been reached by the student.
The Introduction chapter is well structured in introducing the thesis goals and context. Chapter one provides an overview of big data. Chapter two describes the data lake concept. Chapter three is the core part of the thesis developing the decision model, which is then applied to some *"hypothetical"* cases in chapter four. Closing the thesis with conclusions and future works is chapter five.

Strong points. The thesis covers an interesting topic with an engrossing research flow: going from the literature to the search for the main factors, then a survey for refinement, and then application of the decision model. The segmentation of respondents of the survey is maybe more interesting than the decision model itself, as it can give information about what really matters for practitioners. The impression is also that the student spent some considerable effort in the process. Overall the results can be beneficial to raise more awareness about the data lake concept.

Weak points. There are some issues in the conduction of the overall research, so the final outcome could have been much more useful, in my view. The final model is very simplistic in terms of the number of factors considered — it is also completely binary in the decisions at each point *(Y/N)*. Is reality really based on binary decisions? A weighted and parametrized model could have been better if supported by an appropriate research methodology.

- *The literature review* is rather weak: it seems that the factors are collected from a limited set of papers (I count *six* papers on *p.36*). I am not sure some of the factors are really important in the decision, like *"the system that needs to be easy to operate and user friendly"*—not sure the data lake will be impacted by this (in fact, in the revision of the model, this factor is removed). The initial model is rather simple, and is very abstract, with not many parameters to decide. The representation is also a bit misleading, e.g., in the lower part when four boxes are represented, *"yes, also"* means the four conditions are in AND, or in OR? Ending points could have been marked differently, to highlight when the model ends. To note that by removing the ending points, there are not many decision nodes, so they could be even memorized to make a reasoned choice, making the whole idea of a decision model redundant.
- *The survey* is not well reported: what was the response rate and how was the specific target population reached? We know that it is a mixture of academicians, experts in big data, etc... but what are the proportions and from which countries do they come from? Furthermore, there are also factors and numerical evaluations / clustering of respondents, but then in the final model no numerical aspects are taken into account, plus the selection of the factors is not transparent nor based on some objective criteria (the student reports that factors were *hand-picked*).
- *The quality of the text* is in many places sub-par in terms on English writing, and writing style. Just some examples of issues that could have been caught easily by some grammar / spell checker: p.7 *"Unstructured data is are..."*, p.9, *"...we have introduces..."*, p.9, *"Volume, variety, velocity, veracity, value, variability."* (this cannot be a complete sentence) – there are also differences between *"its"* and *"it's"*. References *[6],[7],[41],[46]* are incomplete. Some tables and figures are *"dumped"* in the main text without any in-text comment / reference to them, for example, section 2.6 is just a table (2.1) without any text. Furthermore, the same table 2.1 structure is taken from an online resource: `https://www.guru99.com/data-lake-vs-data-warehouse.html` - some parts are rephrased, but some cells are the same without any indication of the source.

Question(s) for the defence:

- *Over the course of your studies, did you have any hands-on experience with any of the technologies you mention in the thesis (e.g., Hadoop, Spark, Kafka, Flume, etc...)?*
- *Before the creation of the decision model, did you look at previous research proposing some decision model in slightly different domains? e.g., some decision model that can be used for the selection of some cloud-based technology or similar?*
- *Tables 3.2-3.4: what is the meaning of the green highlighting? How were the factors selected? If the values are averages from the Likert values (in the captions is written about a generic "statistic"). Is it good to use averages of Likert values? What are pro and cons?*

Overall, I consider this thesis quite interesting in terms of the research flow. If more accurate in the execution, it could have been an excellent work, but in the current state there are quite some relevant issues. Therefore, I would recommend accepting the work as Master Thesis of the Faculty of Informatics, with the proposed assessment of ***"Satisfactory"* (D)**.

Brno January 16, 2020 Bruno Rossi, Ph.D.