Prof. Esteban Zimányi
Université libre de Bruxelles, Belgium
esteban.zimanyi@ulb.be

August 18, 2021

Reviewer's opinion
on Ph.D. dissertation authored by

Syed Muhammad Fawad Ali

entitled:

Parallelization of User-Defined Functions in an ETL Workflow

1. Problem and its impact

The dissertation of Mr. Syed Muhammad Fawad ALI summarizes research that he has performed as a PhD student enrolled at the Poznan University of Technology under the supervision of Prof. Robert Wrembel.

Mr. ALI's dissertation concerns the field of Extraction, Transformation, and Loading (ETL) workflows, which are relevant in data integration architectures, in particular in data warehouses, data lakes, and data science applications. Nowadays, ETL workflows typically move huge volumes of data between data repositories while executing complex cleaning and data transformation tasks. Consequently, ETL workflows are very time consuming and require to be executed on a cluster to enable parallelization. Furthermore, the variety of data and the wide range of analytical use cases require the ETL developer to develop user-defined functions (UDFs), which put the responsibility on the developer to write correct and efficient ETL code.

2. Contribution

Mr. ALI's dissertation focus on the optimization of computing-intensive UDFs in ETL workflows by means of parallelization. Concretely, Mr. ALI's thesis approaches this problem by:

1. Developing a state-of-the-art analysis of methods and techniques for each stage of the ETL development lifecycle, thus covering conceptual and logical design as well as physical implementation.
2. Developing a state-of-the-art analysis of methodologies for optimizing ETL workflows in each stage of the lifecycle and evaluate them based on a set of selected metrics.

This state-of-the-art analysis allowed Mr. Ali to define the two research challenges addressed in his thesis, that is,

1. Design an ETL framework that separates the production of efficient ETL code from the parallelization concerns
2. Developing a cost model that generates parallelizable UDFs by determining their degree of parallelism to be executed in a parallel distributed environment.

The dissertation is composed of seven chapters and a bibliography.

The first chapter is introductory, it provides a background on the ETL domain, defines the research problems and challenges addressed in the thesis, specifies its contributions, and provides an overview of the thesis and its relationship with the publications of the author.

Chapter 2 is devoted to a state-of-the-art review of ETL workflow design at the conceptual, logical, and physical levels, where the related work on each of these levels is evaluated based on appropriate

metrics proposed by the author. The chapter concludes by summarizing open research and technological issues.
Chapter 3 is devoted to a state-of-the-art review of various ETL optimization techniques. Based on a running example, various optimization techniques are analysed. These techniques include the state-space approach, dependency graph, scheduling strategies, reusable patterns, and parallel strategies, quality metrics, and statistics collection. After describing several commercial ETL tools, the chapter concludes by summarizing the results obtained and pointing to open research and technological issues.
Chapter 4 constitutes a first major contribution of the thesis, that is, the architecture of the proposed extendable ETL framework that addresses the challenges highlighted in the two previous chapters. The proposed framework is composed of four modules as follows. The UDF component assists ETL developers in writing parallelizable UDFs by separating parallelization concerns from the code. The Recommender component includes an extendable set of machine learning algorithms to optimize a given ETL workflow (based on metadata collected during past ETL executions) and to generate a more efficient version of the workflow. The Cost Model component is composed of a library of cost models that can be used by the Recommender to determine the optimal design. Finally, the Monitoring Agent component responsible for monitoring ETL workflow executions, identifying performance bottlenecks, report errors, schedule executions, and gather various performance statistics.
Chapter 5 delves in the UDF component of the ETL framework. By means of a running example used throughout the chapter, it explains the Orchestration Style Sheet (OSS) processor which uses parallel algorithmic skeletons (PAS) to support multiple, potentially differently parallelized target platforms. The chapter continues by describing how OSSs are used for generating parallelizable UDFs. This approach is then illustrated by discussing the use of Map-Reduce OSS for a sentiment analysis use case. The chapter provides an experimental evaluation of the feasibility of the proposed approach and gives general conclusions on the approach.
Chapter 6 tackles the problem of the optimization of UDFs in data-intensive workflows and presents a cost model that determines the degree of parallelism for both case-based and generic parallelizable UDFs. The model is then extended for a machine learning pipeline in order to enable data scientists to choose the best possible machine learning model based on user-defined performance metrics.
The dissertation is concluded in Chapter 7 with a summary of the presented results and indications for further research.

3. Correctness

The main contributions proposed in the thesis are validated by case studies proving the feasibility of the proposed approaches.

4. Knowledge of the candidate

The main contributions of the thesis are based on two very broad state-of-the-art chapters that have been published on a top journal of the domain. The result of this analysis drives the research contributions of the overall thesis.

5. Other remarks[1]

The present reviewer deems the dissertation to be original, accurate, and innovating. In particular, the dissertation contains clearly identified, significant contributions to the field of ETL workflow

[1] Optional

optimization. These contributions are novel, and their experimental evaluation shows their feasibility in practice.

The quality of the research described by Mr. ALI in his dissertation is testified by the publications it has produced:

1. 2 papers in international high-quality journals with impact factor, i.e., 'VLDB Journal' and 'International Journal of Applied Mathematics and Computer Science';
2. 3 papers presented at internationally respected conferences and workshops, i.e., 'Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)', 'European Conference on Advances in Databases and Information Systems (ADBIS)' and 'Big Data Analytics and Knowledge Discovery (DaWaK)'.

With his dissertation, Mr. ALI has demonstrated his skill to work in a very structured and precise manner and competes with colleagues at the highest international standard.

The dissertation is well-written. It clearly states the results, as well as their grounding in the scientific literature. The methodologies used to develop the results follow widely accepted community norms, such as comprehensive literature survey, empirical study, and extensive practical use-case validation.

In conclusion, the present reviewer that the work of Mr. ALI is of high scientific quality, that it is a significant contribution to the field of ETL workflow optimization, and that it contains a large number of significant results.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 13 of *the Act of 14 March 2003 of the Polish Parliament on the Academic Degrees and the Academic Title* (with amendments)[2], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of **ETL Workflows**?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

Signature

[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf