Dr. Oscar Romero
Universitat Politècnica de Catalunya (UPC)
oromero@essi.upc.edu

5 July 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

Extract-Transform-Load (ETL) workflows are a crucial component of Data Warehousing. They are responsible for extracting data from sources, transform and prepare it to then load it into an integrated repository, typically called the data warehouse (DW) that represents a single source of truth for the data analyst. Most of the effort in DW projects boils down to create, maintain and optimize the ETL workflows.

ETL workflows must be tailored to each project. Depending on the quality and nature of each source, one may need to execute a certain number of steps to extract, transform (which includes joining, uniting or performing set operators on such data with that of other sources) and load the data into the DW. Unfortunately, even if there has been lots of efforts to characterize the required operators needed in the ETL workflows, there is a strong consensus, both in academia and professional practitioners, around the need to consider user-defined functions (UDFs). UDFs are, basically, the way we have to tailor and personalize ETLs to our very own needs. UDFs are defined as an arbitrary code generated by the responsible data steward designing the ETLs and provide the required flexibility and adaptability to different realities and domains.

Importantly, the ETL performance optimization is a relevant aspect, which, given its complexity, has been traditionally overlooked. Nevertheless, (i) current approaches focus on relational algebraic-ish operators and reordering to optimize its scheduling (provided the semantics of the operators are crystal clear). This has been, traditionally, the approach used by academic prototypes to optimize ETLs. (ii) Another approach is by scaling up (either vertically or horizontally). Interestingly, this approach is more extended in industry and we can find different tools following this idea and parallelizing ETL workflows to some extent.

However, there is not a single solution considering UDFs neither in (i) nor (ii). To my knowledge, this thesis is the very first one considering such relevant aspect of ETLs.

Las, but not least, this work has gained even more impact with the arrival of Big Data and Data Science. The concept of ETL, typically related to Data Warehousing, has been broadened and generalized to consider any data flow between two repositories. In the Big Data field, they are known as data workflows, but in essence, they are generalized ETLs. By generalized I mean data flows that can be easily tailored and adapted to the current needs (e.g., big volumes of data, large variety of formats and sources, etc.). As such, UDFs are essential in the Big Data realm and provide the required

flexibility to generalize the concept of ETL to data workflow. Thus, UDFs are even more relevant in the Big Data context and optimizing them is crucial and therefore worth to research.

Consequently, the impact of the contributions presented in this thesis have high impact, both for the academia and industry.

2. Contribution

This thesis focuses on a specific kind of ETL optimization: the parallelization of computer-intensive UDFs in ETL workflows.

The thesis objective is ambitious, since UDFs are black-boxes (directly provided in a programmatic manner: e.g., in Java, Python, etc.) for the system that they typically execute without interpreting or understanding its precise semantics. Therefore, in order to optimize UDFs they must be defined in a way the system can understand and optimize them (in this case, optimize means parallelize). As any other topic not properly tackled before, the problem lacks formalization and therefore, the first challenge is to create a comprehensive framework capturing the needed components to attain the objective set.

The thesis defines two main objectives according to the discussion above:

1. The first challenge is to design a comprehensive framework to facilitate the design of UDFs that can be later be interpreted and optimized. The current proposal highlights two aspects that this framework must achieve:
    a. Identify the right parts of the UDF that can be parallelized and those that cannot. This is challenging. On the one hand, this is important because automatically parallelizing code not prepared to be parallelized may introduce errors in the execution. On the other hand, this objective should prevent the ETL designer from writing parallelizable code, since this is well-known to be error-prone, hard and requiring an expertise most ETL developers do not have.
    b. For those aspects that might be parallelized, determine the degree of parallelism achievable.
2. Next, a solution is proposed by instantiating the two main components of the framework proposed. First, a component to guarantee efficient and parallelizable UDFs and then a cost model meant to choose the right configuration of the parallelizable version of a UDF.

Importantly, being a previously poorly explored problem, this thesis required a huge effort understanding the potential related work and considering it to propose the abovementioned framework (objective 1) and a solution (objective 2). This thesis excels in this aspect and, indeed, the strongest publication (the VLDB J) shows strong foundations based on a rigorous and thorough state of the art, on which the reference framework and the provided solution build on top. Therefore, the research methodology of this thesis is excellent and grounded on the state of the art.

Last, but not least, this thesis presents a prototype following the ideas presented in the thesis. This is an excellent way to wrap up the discussion and show the practical impact of the contributions proposed.

The author makes a good summary of the thesis contributions, motivates and contextualizes them and grounds discussions in a thorough literature review. The validation in the form of a prototype

highlights the high impact of the results of this thesis that, importantly, open many other research paths from the overall framework presented.

Aside, I would like to refer to the following additional quality indicators.
First, the impact of the publications. Below the list follows:

- S.M.F. Ali, R. Wrembel: Framework to Optimize Data Processing Pipelines Using Performance Metrics. Int. Conf. on Big Data Analytics and Knowledge Discovery (DaWaK), pp. 131-140, LNCS 12393, Springer, 2020.
- S.M.F. Ali, J. Mey, M. Thiele: Parallelizing user-defined functions in the ETL workflow using orchestration style sheets. Int. Journal of Applied Mathematics and Computer Science (AMCS), 29:(1), pp. 69-79.
- S.M.F. Ali, R. Wrembel: Towards a Cost Model to Optimize User-Defined Functions in an ETL Workflow Based on User-Defined Performance Metrics. European Conf. on Advances in Databases and Information Systems (ADBIS), pp. 441-456, LNCS 11695, Springer 2019.
- S.M.F. Ali: Next-generation ETL Framework to Address the Challenges Posed by Big Data. Int. Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) @EDBT/ICDT, CEUR-WS, vol. 2062, 2018.
- S.M.F. Ali, R. Wrembel: From conceptual design to performance optimization of ETL workflows: current state of research and open problems. The VLDB Journal, 26:(6), pp. 777-801, 2017.

Mr. Ali published a fair amount of papers during his PhD and all of them either in well-known international conferences (DaWaK, DOLAP and ADBIS) or top journals (AMCS, VLDB J). It is fair to highlight the VLDB J publication, which one of the best venues Mr. Ali could have submitted to.
Other quality indicators follow:

- Achieving the minimum requirements to defend in two top universities (Poznan and Dresden),
- A fair amount of citations (~50) according to Google Scholar,
- A solution to an extremely difficult problem for which there is no real alternative neither in the market nor in the scientific literature.

3. Correctness

The first step to guarantee correctness is to ground your work on the state of the art. For this topic, however, there was nothing such as a previous literature review. Chapter two presents an outstanding literature review on ETL modelling that has a huge impact given it is the very first of its kind.
This Chapter provides a clear, comprehensive story-telling facilitating to the readers their journey into the topic. First, a comprehensive discussion about conceptual and logical ETL modelling takes place. Then, how to transform them into specific implementations follows. All these analyses are done by comparing advantages and disadvantages of each solution.
To my knowledge, this Chapter covers all different manners I can think of to model ETLs and makes an excellent work comparing them with pros and cons. As any relevant literature review, this Chapters ends with a list of relevant open issues.

If Chapter 2 had, per se, high added value to the research community, Chapter 3 does an excellent job characterizing the current trends to optimize ETL workflows. This topic is even less researched and more disperse than ETL modelling and, in my opinion, these two facts alone make Chapter 3 more relevant. The difficulty of this Chapter is that many relevant works might not be easy to spot (for the wording or application used) and also because industry implemented certain techniques that require to

analyse and understand how their software works. The author proposes a set of metrics and a running example that facilitate the reading. Following the structure of Chapter 2, it presents a nice summary of pros and cons. This Chapter is outstanding and I appreciate it very much to contextualise and frame the thesis properly. Its value and impact for the research community and ETL practitioners is very high.

Chapter 4 then presents the "Next-Gen ETL Framework". This Chapter reads very natural after the previous Chapters. It presents a framework to design ETL workflows and it is the main contribution of the thesis and a natural consequence of the good work done in Chapters 2 and 3. As previously said, this Chapter is a must after understanding in detail the current state of the problem at hand. Thus, the author presents a clear framework that organises and separate concerns and, as such, facilitates conducting research in this problem. This framework builds on top of an ETL tool and spans four components: the UDFs component, the recommender, the cost model and the monitoring agent. The resulting parallelizable UDFs are executed in a component called the distributed framework. The Chapter sits at a conceptual level with a proper definition of the tasks and concerns every module covers and therefore, representing the needed conceptualization of the problem.

Reasonably, the thesis does not tackle all the components discussed in the conceptual framework and focuses on two of them (Chapter 5 and 6 respectively). Chapter 5 maps to the UDFs component of the framework and Chapter 6 to the cost model.
Chapter 5 is clear, nicely written and based on aspects previously surveyed in Chapter 3. It is a pragmatic and feasible view of the problem and it reads sound and rigorous. The experiments presented in this Chapter are reasonable, given this is a first attempt to fill the gap in this aspect. Focused on execution time and lines of code written, they show the feasibility of the module proposed. This is the first work towards an out-of-the-box functionality to write parallel UDFs. This work opens the door to even higher-level approaches based on the findings presented here.
Chapter 6, in turn, presents a cost model to identify the best configuration for UDFs. A cost-model to optimize black-box-ish code and selecting its optimal configuration of parallelizable UDFs is a very relevant contribution. This Chapter also presents the idea to consider machine learning models as part of the solution. Combining a classical approach, i.e., a cost-model, with innovative predictive algorithms is novel and worth to consider, as discussed in the experiments.

I appreciate all solutions proposed are extensible and easily adaptable. This is sound since, in the long term, the main contribution is the framework in Chapter 4, and Chapter 5 and Chapter 6 are meant to show the feasibility of the tasks assigned to the two main components. From the work done in Chapters 5 and 6, one may come up with plenty of innovative ideas resulting from a thorough study of a, previous to this thesis, ill-defined and poorly explored topic.

I would not like to finish without highlighting the huge amount of work done to show the feasibility of the framework presented. All techniques and methods are technically sound, well-grounded on the excellent literature review in Chapters 2 and 3 and framed with an innovative and excellent contribution in the form of a framework (Chapter 4). The results of the thesis are excellent.

4. Knowledge of the candidate

The story-telling in this thesis is outstanding. The general knowledge of the author in the field of UDFs within ETL workflows is very high and makes him, without a doubt, one of the most informed

persons in the field. This allowed him to write a thesis that is very easy to read and introduces the reader nicely into a very hard topic.

Specifically, as discussed in the previous section of this document, this is shown by his ability to conduct an outstanding literature review that grounds the rest of his research. The thesis is very well-written, is comprehensive, and denotes a deep understanding of the topic. Chapter 4 presents a framework in a convincing way, as consequence of the results presented in Chapters 2 and 3. Chapter 5 and 6 show the feasibility of the two main modules of the framework proposed and overall, removes any big concern the reader may have when considering the framework proposed.

Mr. Ali has shown he masters the topic, he is able to perform top-quality research in the field of Information and Communication Technology, communicate his results in an excellent manner with a well-structured, clear and concise document, and showcase its viability with realistic prototypes. The methodology of the thesis is sound, with a clear state-of-the-art, followed by the design of a solution and its experimentation and validation. All this is wrapped-up with top quality publications in well-reputed forums in the area of Information and Communication Technology.

I have no hesitation the knowledge of the candidate is more than enough to guarantee the PhD.

5. Other remarks[1]

This thesis was conducted in the frame of an international PhD programme (IT4BI-DC). As result, Mr. Ali had to spend time abroad (Univ. Dresden). This is a relevant aspect since internationalization is a must in research. Conducting a long stay in Germany was for sure relevant since Mr. Ali could learn the pros and cons of conducting research in two different countries. This is for sure a valuable asset for a researcher, even if it also adds complexity to the setting since changing context and the mandatory mobility might not be easy from the personal point of view.

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 ....?

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

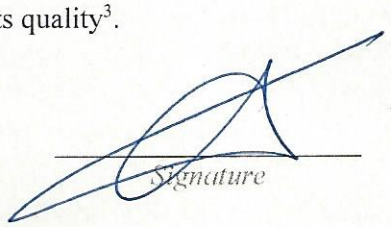
[1] Optional

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

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*

Therefore, I **recommend to distinguish** the dissertation for its quality[3].



[3] Obviously, this sentence is optional.