PARALLELIZATION OF USER-DEFINED ETL TASKS IN AN ETL WORK FLOW

Syed M. Fawad Ali

syed.s.ali@doctorate.put.poznan.pl

Supervisors

Prof. Robert Wrembel – PUT

Prof. Wolfgang Lehner – TUD

1. Background

The minimization of the execution time of an ETL workflow is of particular importance, since ETL workflows have to complete their task within a specific time window.

Open Issues

- **Automatic optimization techniques for ETL workflows**
- **Monitoring system that identifies bottlenecks and gives suggestion to improve ETL workflow performance**

2. Objective

- **Design a framework to exploit parallelization for user-defined tasks in an ETL workflow and to enhance its execution performance.**
- **Develop a cost function to check whether it is feasible to exploit parallelization in a particular scenario or not.**

3. Methodology

Phase I Study on different skeletons	• Conceptual Study to answer the following questions: • What is the semantic of Parallel Algorithmic Skeletons (PAS)? • What PAS are applicable or compatible with ETL workflows? • How can we use one PAS with other PAS?
Phase II Feasibility of the Approach	• Cost function to answer the following questions: • Whether it is feasible to parallelize the ETL workflow or not? • Whether partition scheme is compatible or not? • Where and when we should split pipelines, maintain pipelines, merge pipelines or partially merge pipelines? • When does it make sense to exploit parallelism?
Phase III Degree of Parallelization	• Check to which degree we should implement parallelization.

4. Ongoing Work

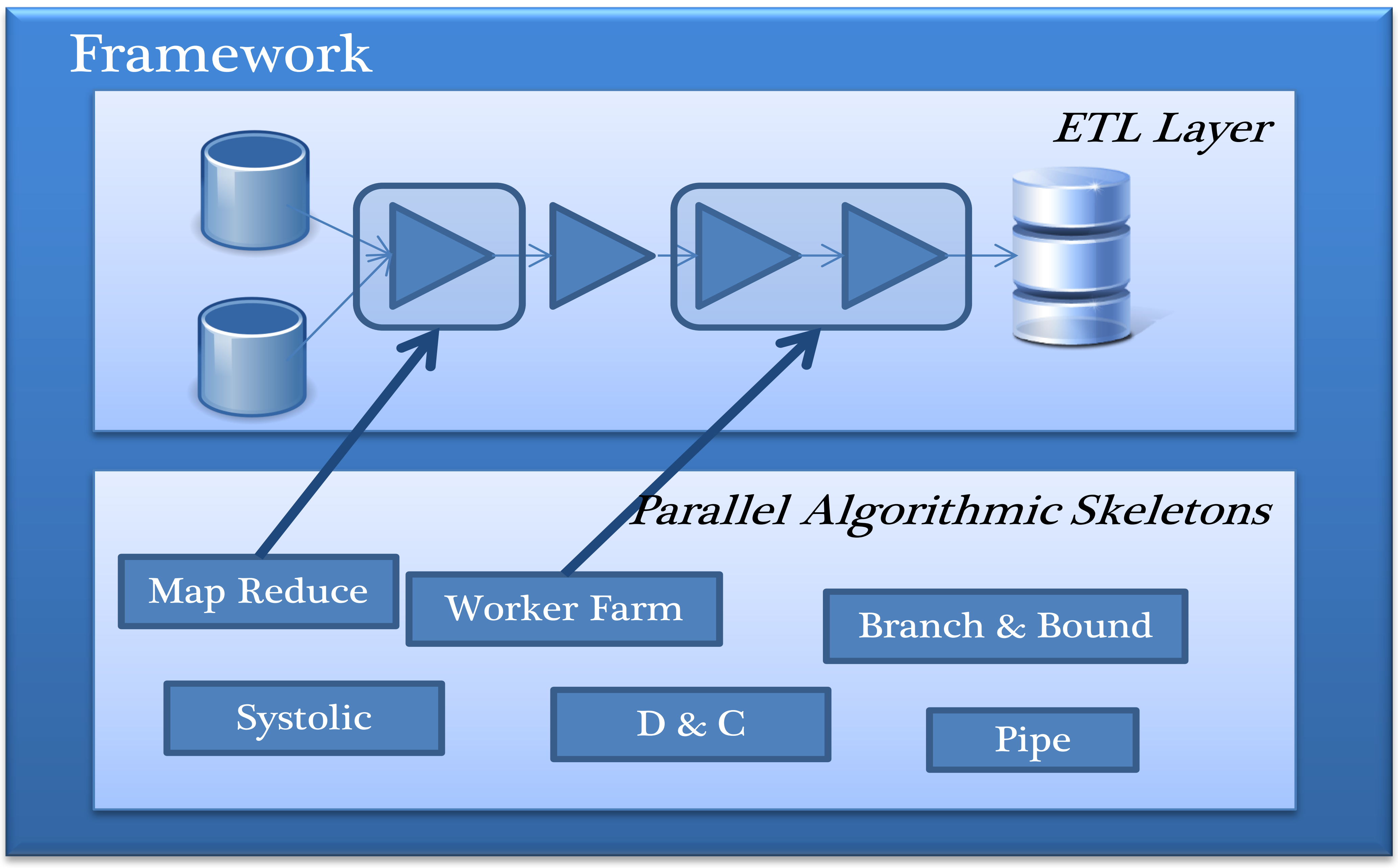
- Checking compatibility and applicability of different PAS on ETL workflow using "Pentaho Data Integrator" as a sandbox.
- Systematic Literature Review on ETL performance optimization.

5. Conclusion

Final Product

A framework which suggests the ETL developer to choose a sequence of skeletons from the extensible library of PAS's defined in our framework.

The sequence of skeletons suggested by the framework will enable the ETL developer to exploit parallelism in an ETL workflow to achieve better performance.



References

ORCA - Open paRallel Computing Architecture - Bornhovd, Lehner 2013

Easy & Effective Parallel Programmable ETL - Thomsen, Pedersen 2011

State Space Optimization of ETL Workflows - Simistis, Vassiliadis, Sellis 2005

Generic & Customizable Framework for ETL Scenarios - Vassiliadis, Simistis, Georgantas, Terrovitis & Skiadopoulos 2005