Efficient Execution of Scientific Workflows in the Cloud through Adaptive Caching

G. Heidsieck[1], D. de Oliveira[3], E. Pacitti[1], C. Pradal[2], F. Tardieu[4], P. Valduriez[1]

[1] *Inria, LIRMM, Université de Montpellier, Montpellier, France,*
[2] *AGAP, CIRAD, Montpellier SupAgro, France,*
[3] *Institute of Computing, UFF, Rio de Janeiro, Brazil,*
[4] *INRA, LEPSE, Montpellier, France.*

Key words — Adaptive Caching, Scientific Workflow, Cloud, Workflow Execution.

Abstract :

In many scientific domains, *e.g.*, bio-science [1], complex experiments typically require many processing or analysis steps over huge quantities of data. They can be represented as scientific workflows (SWfs), which facilitate the modeling, management and execution of computational activities linked by data dependencies. As the size of the data processed and the complexity of the computation keep increasing, these SWfs become data-intensive [1], thus requiring execution in a high-performance distributed and parallel environment, *e.g.* a large-scale virtual cluster in the cloud [2].

It is common for workflow users to reuse other workflows or data generated by other workflows. Reusing and re-purposing workflows allow for the user to develop new analyses faster [3]. Furthermore, a user may need to execute a workflow many times with different sets of parameters and input data to analyze the impact of some experimental step, represented as a workflow fragment, *i.e.* a subset of the workflow activities and dependencies. In both cases, some fragments of the workflow will be executed many times, which can be highly resource consuming and unnecessary long. Workflow re-execution can be avoided by storing the intermediate results of these workflow fragments and reuse them in later executions.

In a single user perspective, the reuse of the previous results can be done by storing the relevant outputs of intermediate activities (intermediate data) within the workflow. This requires the user to manually manage the caching of the results that she wants to reuse, which can be difficult as she needs to be aware of the data size, execution time of each task, *i.e.* the instantiation of an activity during the execution of a workflow, or other factors that could allow deciding which data is the best to store.

The solution proposed is an adaptive caching solution for efficient execution of data-intensive SWfs in the cloud [4]. By adapting to the variations in tasks' execution times, our solution can maximize the reuse of intermediate data produced by SWfs from multiple users. Our solution is based on a new SWfMS architecture that automatically manages the storage and reuse of intermediate data.

References

[1] Pradal, C., Cohen-Boulakia, S., Heidsieck, G., Pacitti, E., Tardieu, F., & Valduriez, P. (2018). Distributed management of scientific workflows for high-throughput plant phenotyping. ERCIM News, (113), 36-37.

[2] Liu, J., Pacitti, E., Valduriez, P., & Mattoso, M. (2015). A survey of data-intensive scientific workflow management. Journal of Grid Computing, 13(4), 457-493.

[3] Pradal, C., Artzet, S., Chopard, J., Dupuis, D., Fournier, C., Mielewczik, M., Negre, V., Neveu, P., Parigot, D., Valduriez, P. & Cohen-Boulakia, S. (2017). InfraPhenoGrid: a scientific workflow infrastructure for plant phenomics on the grid. Future Generation Computer Systems, 67, 341-353.

[4] Heidsieck, G., de Oliveira, D., Pacitti, E., Pradal, C., Tardieu, F., & Valduriez, P. (2019) Adaptive Caching for Data-Intensive Scientific Workflows in the Cloud. Int. Conf. on Databases and Expert Systems Applications.