Tracking of Online Parameter Fine-tuning in Scientific Workflows

Renan Souza, Vitor Silva, José Camata, Alvaro L G A Coutinho, Patrick Valduriez, Marta Mattoso

Tracking of Online Parameter Fine-tuning in Scientific Workflows

Renan Souza[§,°], Vítor Silva[§], Jose J. Camata[§]
Alvaro L. G. A. Coutinho[§], Patrick Valduriez[¶], Marta Mattoso[§]
[§]COPPE/Federal University of Rio de Janeiro, [°]IBM Research Brazil, [¶]Inria and LIRMM, Montpellier

EXTENDED ABSTRACT

In typical large-scale scientific applications, several parameters of complex computational models have to be predefined in a simulation, each with a wide range of possible values. Listing all possible combinations of parameters and exhaustively trying them all is nearly impossible even in extreme-scale High Performance Computing (HPC). There may be a huge number of possible combinations and processing each one may take several hours or days, making the whole computation last for weeks or months. Typically, after the initial set ups, the scientist starts the computation and occasionally fine-tunes specific parameters based on intermediate result analysis. The term "human-in-the-loop" is used when computational scientists can actively participate in the computational process. Specific adaptations can generate an important improvement on performance, resource consumption, and quality of results [2]. To allow for online human adaptation, dynamic workflow solutions are required. Most existing workflow solutions do not allow for online human adaptations, which is considered a future research challenge in a recent survey [4]. Chiron[3], WorkWays [7], and Copernicus [8] are a few exceptions that allow for online data adaptation.

Parameter tuning – the subject of research of this work – is only one among many other types of adaptations possible in human-adapted workflows in HPC [6]. Registering the adaptations is essential to track and analyze the effects of fine-tunings. In [1], the authors discuss past, present and future of scientific workflows, and as a future issue they mention that "*monitoring and logging will be enhanced with more interactive components for intermediate stages of active workflows.*" We did not find any work that registers workflow adaption in logs or in provenance databases. This work aims at capturing and registering such human adaptation data (*e.g.*, values before and after a specific parameter fine-tuning; reason for the tuning), relating them with other relevant data (*e.g.*, domain-specific strategic values and execution data), and allowing all these data to be efficiently integrated. This contributes for online data analysis and data-driven decisions (*e.g.*, how a specific user action impacted the processing time), helping to put humans in the online loop of large-scale scientific computing. Also, recording those adaptations contributes to the results' reliability and reproducibility.

We developed DfAdapter [5], a tool that collects human adaptations in the dataflow, while the workflow runs. It controls and stores specific parameter-tunings in a provenance database, relating the human adaptation data with data for: domain, dataflow provenance, execution, and performance.

As shown in Figure 1, initially DfAdapter registers the user Bob, who is going to adapt the dataflow; then it registers identifiers of the current state of the workflow (*e.g.* step *i* of the loop). To track the tunings, it receives from Bob the set of parameters, *e.g. attr5* to be modified to "*val5*" into *Dataset2*. Then, DfAdapter modifies the values in *Dataset2*. Finally, it registers the iteration counter, the execution state at the adaptation moment, the dataset, values before and after, and the current wall-time all in the provenance database. Relevant insight is obtained with visualizations complemented by tracking queries like: "List all Bob's tunings correlating with time step" or "Avg. of values 10 iterations before and after the tunings". DfAdapter can be coupled to a workflow managed by a parallel workflow management system or by a workflow defined using an HPC library, or even a script. DfAdapter works as a debugging tool on an instrumented code.

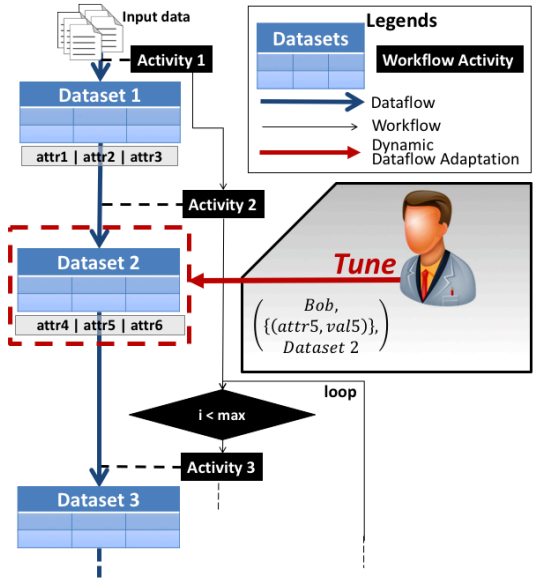

Figure 1. Tuning parameters in a dataset in a dataflow.

Acknowlegments. This work was partially funded by CAPES, CNPq, FAPERJ and Inria (MUSIC and SciDISC projects), EU HPC4E H2020 Programme and MCTI/RNP-Brazil

REFERENCES

[1] Atkinson, M., Gesing, S., Montagnat, J., Taylor, I. Scientific workflows: past, present and future. *FGCS*, 75:216–227, 2017.
[2] Deelman, E., Peterka, T., Altintas, I., Carothers, C.D., Kleese van Dam, K., Moreland, K., Parashar, M., Ramakrishnan, L., Taufer, M., et al. The future of scientific workflows. *Int. Journal of HPC Applications*, 2017.
[3] Dias, J., Guerra, G., Rochinha, F., Coutinho, A.L.G.A., Valduriez, P., Mattoso, M. Data-centric iteration in dynamic workflows. *FGCS*, 46(C):114–126, 2015.
[4] F. da Silva, R., Filgueira, R., Pietri, I., Jiang, M., Sakellariou, R., Deelman, E. A characterization of workflow management systems for extreme-scale applications. *FGCS*, 75:228–238, 2017.
[5] GitHub. DfAdapter Repository. Available at: https://github.com/hpcdb/DfAdapter
[6] Mattoso, M., Dias, J., Ocaña, K.A.C.S., Ogasawara, E., Costa, F., Horta, F., Silva, V., de Oliveira, D. Dynamic steering of HPC scientific workflows: A survey. *FGCS*, 46:100–113, 2015.
[7] Nguyen, H.A., Abramson, D., Kiporous, T., Janke, A., Galloway, G. WorkWays: interacting with scientific workflows. *Gateway Computing Environments Workshop*, 21–24, 2014.
[8] Pouya, I., Pronk, S., Lundborg, M., Lindahl, E. Copernicus, a hybrid dataflow and peer-to-peer scientific computing platform for efficient large-scale ensemble sampling. *FGCS*, 71:18–31, 2017.