Geophysical Research Abstracts
Vol. 18, EGU2016-14793-1, 2016
EGU General Assembly 2016

Visualisation methods for large provenance collections in data-intensive collaborative platforms

Alessandro Spinuso (1), Rosa Fligueira (2), Malcolm Atkinson (2), and Andre Gemuend (3)

(1) KNMI, R&D Observation and Data Technology, Utrecht, The Netherlands, (2) University of Edinburgh, School of Informatics, Edinburgh, UK, (3) Fraunhofer SCAI, Sankt Augustin, Germany

This work investigates improving the methods of visually representing provenance information in the context of modern data-driven scientific research. It explores scenarios where data-intensive workflows systems are serving communities of researchers within collaborative environments, supporting the sharing of data and methods, and offering a variety of computation facilities, including HPC, HTC and Cloud.

It focuses on the exploration of big-data visualization techniques aiming at producing comprehensive and interactive views on top of large and heterogeneous provenance data. The same approach is applicable to control-flow and data-flow workflows or to combinations of the two. This flexibility is achieved using the W3C-PROV recommendation as a reference model, especially its workflow oriented profiles such as D-PROV (Messier et al. 2013). Our implementation is based on the provenance records produced by the dispel4py data-intensive processing library (Filgueira et al. 2015).

dispel4py is an open-source Python framework for describing abstract stream-based workflows for distributed data-intensive applications, developed during the VERCE project. dispel4py enables scientists to develop their scientific methods and applications on their laptop and then run them at scale on a wide range of e-Infrastructures (Cloud, Cluster, etc.) without making changes. Users can therefore focus on designing their workflows at an abstract level, describing actions, input and output streams, and how they are connected. The dispel4py system then maps these descriptions to the enactment platforms, such as MPI, Storm, multiprocessing. It provides a mechanism which allows users to determine the provenance information to be collected and to analyze it at runtime.

For this work we consider alternative visualisation methods for provenance data, from infinite lists and localised interactive graphs, to radial-views. The latter technique has been positively explored in many fields, from text data visualisation to genomics and social networking analysis. Its adoption for provenance has been presented in literature (Borkin et al. 2013) in the context of parent-child relationships across processes, constructed from control-flow information. Computer graphics research has focused on the advantage of this radial distribution of interlinked information and on ways to improve the visual efficiency and tunability of such representations, like the Hierarchical Edge Bundles visualisation method, (Holten et al. 2006), which aims at reducing visual clutter of highly connected structures via the generation of bundles.

Our approach explores the potential of the combination of these methods. It serves environments where the size of the provenance collection, coupled with the diversity of the infrastructures and the domain metadata, make the extrapolation of usage trends extremely challenging. Applications of such visualisation systems can engage groups of scientists, data providers and computational engineers, by serving visual snapshots that highlight relationships between an item and its connected processes.

We will present examples of comprehensive views on the distribution of processing and data transfers during a workflow's execution in HPC, as well as cross workflows interactions and internal dynamics. The latter in the context of faceted searches on domain metadata values-range. These are obtained from the analysis of real provenance data generated by the processing of seismic traces performed through the VERCE platform.