

Scientific Workflows: Cyberinfrastructure for e-Science

Bertram LUDAESCHER

Department of Computer Science at University of California, Davis

ludaesch@ucdavis.edu

Scientific workflows are the domain scientist's way to harness cyberinfrastructure for e-Science. Through various collaborative projects over the last couple of years, we have gained first-hand experience in the challenges faced when trying to realize the vision of scientific workflows. Domain scientists are often interested in "end-to-end" frameworks which include data acquisition, transformation, analysis, visualization, and other steps. While there is no lack of technologies and standards to choose from, a simple, unified framework combining data and process-oriented modeling and design for scientific workflows has yet to emerge. We highlight the requirements and design challenges typical of many scientific workflows: Raw and derived data products come in many forms and from different sources, including custom scripts and specialized packages, e.g. for statistical analysis or data mining. Not surprisingly, the process integration problems are not solved by "making everything a web service", nor are the data integration problems solved by "making everything XML". The real workflow challenges are more intricate and will not go away by the adoption of any easy, one-size-fits-all solution or standard. The problems are further compounded by the scientists' need to compare results from multiple workflows runs, employing various alternative (and often brand-new) analysis methods, algorithms, and parameter settings. We describe ongoing work to combine various concepts (e.g. semantic types for workflow components, models of computation, data and workflow provenance) and techniques (e.g. actor- and flow-oriented programming) into a coherent overall framework for collection-oriented scientific workflow modeling and design. The initial focus of our work is not on optimizing machine performance (e.g., CPU cycles or memory resources), but on optimizing a more precious resource in scientific data management and analysis: human time.