Chapter 8
Data Management in Scientific Workflows

Ewa Deelman
University of Southern California, USA

Ann Chervenak
University of Southern California, USA

ABSTRACT

Scientific applications such as those in astronomy, earthquake science, gravitational-wave physics, and others have embraced workflow technologies to do large-scale science. Workflows enable researchers to collaboratively design, manage, and obtain results that involve hundreds of thousands of steps, access terabytes of data, and generate similar amounts of intermediate and final data products. Although workflow systems are able to facilitate the automated generation of data products, many issues still remain to be addressed. These issues exist in different forms in the workflow lifecycle. This chapter describes a workflow lifecycle as consisting of a workflow generation phase where the analysis is defined, the workflow planning phase where resources needed for execution are selected, the workflow execution part, where the actual computations take place, and the result, metadata, and provenance storing phase. The authors discuss the issues related to data management at each step of the workflow cycle. They describe challenge problems and illustrate them in the context of real-life applications. They discuss the challenges, possible solutions, and open issues faced when mapping and executing large-scale workflows on current cyberinfrastructure. They particularly emphasize the issues related to the management of data throughout the workflow lifecycle.

DOI: 10.4018/978-1-61520-971-2.ch008

INTRODUCTION

Scientific applications such as those in astronomy, earthquake science, gravitational-wave physics, and others have embraced workflow technologies to do large-scale science (Taylor, et al. editors, 2006). Workflows enable researchers to collaboratively design, manage, and obtain results that involve hundreds of thousands of steps, access terabytes of data, and generate similar amounts of intermediate and final data products. Although workflow systems are able to facilitate the automated generation of data products, many issues still remain to be addressed (Gil, Deelman et al. 2007). These issues exist in different forms in the *workflow lifecycle* (Deelman and Gil 2006). We describe the workflow lifecycle as consisting of a workflow generation phase where the analysis is defined, the workflow planning phase where resources needed for execution are selected, the workflow execution part, where the actual computations take place, and the result, metadata, and provenance storing phase.

During workflow creation, appropriate input data and workflow components need to be discovered. During workflow mapping and execution, data need to be staged-in and staged-out of the computational resources. As data are produced, they need to be archived with enough metadata and provenance information so that they can be interpreted and shared among collaborators. This chapter describes the workflow lifecycle and discusses the issues related to data management at each step. We describe challenge problems and, where possible, illustrate them in the context of the following applications: the Southern California Earthquake Center (SCEC) CyberShake (Maechling, Chalupsky et al. 2005), an earthquake science computational platform; Montage (Berriman, Deelman et al. 2004), an astronomy application; and the Laser Interferometer Gravitational Wave Observatory's (LIGO) binary inspiral search (Brown, Brady et al. 2006), a gravitational-wave physics application. These computations, represented as workflows, are running on today's national cyberinfrastructure and use workflow technologies such as Pegasus (Deelman, Mehta et al. 2006) and DAGMan (Couvares, Kosar et al. 2006) to map high-level workflow descriptions onto the available resources and execute the resulting computations. This chapter describes the challenges, possible solutions, and open issues faced when mapping and executing large-scale workflows on current cyberinfrastructure. We particularly emphasize the issues related to the management of data throughout the workflow lifecycle. In addition to presenting these issues, we describe particular solutions and existing approaches to the problem.

WORKFLOW CREATION

From the point of view of data, the workflow lifecycle includes the following transformations (see Figure 1): data discovery, setting up the data processing pipeline, generation of derived data, and archiving of derived data and its provenance. Data analysis is often a collaborative process or is conducted within the context of a scientific collaboration. An example of such a large-scale collaboration is the LIGO scientific Collaboration (LSC), which brings together physicists from around the world in a joint effort to detect gravitational waves emitted by celestial objects (Barish and Weiss 1999). In astronomy, projects such as Montage develop community-wide image services. In earthquake science, scientists bring together community models to understand complex wave propagation phenomena.

Data and Software Discovery

Scientists in a collaboration frequently submit workflows to process data sets and derive scientific knowledge. These collaborators may submit related workflows and build upon earlier work by other scientists. Thus, scientists need to be able to