Towards Next Generation Provenance Systems for e-Science

Fakhri Alam Khan, University of Vienna, Austria

Sardar Hussain, University of Glasgow, UK

Ivan Janciak, University of Vienna, Austria

Peter Brezany, University of Vienna, Austria

ABSTRACT

e-Science helps scientists to automate scientific discovery processes and experiments, and promote collaboration across organizational boundaries and disciplines. These experiments involve data discovery, knowledge discovery, integration, linking, and analysis through different software tools and activities. Scientific workflow is one technique through which such activities and processes can be interlinked, automated, and ultimately shared amongst the collaborating scientists. Workflows are realized by the workflow enactment engine, which interprets the process definition and interacts with the workflow participants. Since workflows are typically executed on a shared and distributed infrastructure, the information on the workflow activities, data processed, and results generated (also known as provenance), needs to be recorded in order to be reproduced and reused. A range of solutions and techniques have been suggested for the provenance of data collection and analysis; however, these are predominantly workflow enactment engine and domain dependent. This paper includes taxonomy of existing provenance techniques and a novel solution named VePS (The Vienna e-Science Provenance System) for e-Science provenance collection.

Keywords: *e-Science, Middleware, Provenance, Workflow, Workflow Enactment Engine*

INTRODUCTION

The main theme of e-Science (Schroeder, 2008) is to promote collaboration amongst researchers across their organizational boundaries and disciplines - to reduce coupleness and dependencies and encourage modular, distributed, and independent systems. This has resulted in dry-lab experiments also known as in-silico experiments (Cavalcanti et al., 2005). Unlike wet-lab experiments, the dry-lab experiments enable a researcher to plan an experiment, locate suitable activities via resource directories, combine them into a workflow, and execute it. e-Science workflows (Taylor et al., 2006) are used to specify the execution order of tasks (i.e.

DOI: 10.4018/jismd.2011070102

activities). A task may take data input, process it, and produce data output. Real world workflows are complex in nature and may contain several hundreds of activities. Scientists need their experimental activities to be recorded in order to be re-usable and re-producible, similar to the used annotation and book logging in wet-lab experiments. Workflow provenance (Khan et al., 2008) describes the workflow service invocations during its execution, information about services, input data, and data produced to help keeping track of workflow activities (Simmhan et al., 2005). It gives not only insight into the workflows, but enables re-execution of workflows as well. Provenance of workflows includes information about the underlying infrastructure, input and output of workflow activities, their transformations, and context used. e-Science workflows are typically executed on a distributed and dynamic infrastructure provided by different institutions - i.e. resources may join and leave continuously. Therefore, provenance, metadata, and annotations of workflows are of paramount importance for reliable and trustworthy e-Science workflows. There is a strong need to propose and build a provenance system that is in-line with the e-Science core theme of modularity and de-coupleness, which ultimately means domain and application independent provenance system. Key requirements for e-Science provenance systems are interoperability, domain independence, light weight, visualization, and report generation. Interoperability means that an e-Science provenance system should readily work across different domains, applications, and workflow enactment engines.

However, the existing research and development work is mainly focused on provenance collection tightly coupled with the workflow enactment engines, often specific to their projects. With the growing e-Science infrastructures there is a strong need for a provenance system that works across multiple domains and enactment engines. We call such a system loosely coupled provenance system. Not only portability is an important issue to address, but also the performance impact of the provenance collection process on the overall infrastructure as well, as provenance collection is an additional task to the core computational processing in e-Science workflows so that it should be lightweight.

The major contribution of this paper is twofold. First, various possible ways and scenarios through which provenance can be collected are discussed. Taxonomy of existing work according to those scenarios is elaborated based on the coupling of the provenance system to a concrete workflow enactment engine. Secondly, the Vienna e-Science Provenance System (VePS) focusing on workflow enactment engine independence, domain independence, portability, and less performance overhead is introduced together with its design, architecture, and the performance evaluation of our prototype implementation.

The rest of the paper is organized as follows. First, the concepts and terminologies used in our approach are introduced, and then the taxonomy of existing solutions for a provenance system is discussed. Introduction to the VePS architecture, design, and implementation is provided. Next we detail and share performance evaluation, experiences, and observed issues. Finally, we conclude our work and outline future development directions.

CONCEPTS AND TERMINOLOGY

e-Science is a science or research theme that exploits Grid- or Cloud-based solutions more often called e-Infrastructure. The term e-Infrastructure is used for the technology that supports research undertaken comprising of distributed and on-demand computing software. e-Science provides researchers with shared access to large data collections, advanced ICT tools for data analysis, large scale computing resources, and high performance visualization, among other examples. According to Greenwood et al. (2003) ``*e-Science is the use of electronic resource instruments, sensors, databases, computational methods, and computers by scientists working*