Chapter 19
Supporting Data-Intensive Analysis Processes: A Review of Enabling Technologies and Trends

Lawrence Yao
University of New South Wales, Australia

Fethi A. Rabhi
University of New South Wales, Australia

Maurice Peat
University of Sydney, Australia

ABSTRACT

Research scientists in data-intensive science use a variety of scientific software applications to support their analyses and processes. To efficiently support the work of these scientists, software applications should satisfy the essential requirements of interoperability, integration, automation, reproducibility, and efficient data handling. Various enabling technologies including workflow, service, and portal can be used to address these essential requirements. Through an in-depth review, this chapter illustrates that no one technology can address all of the essential requirements of scientific processes and therefore necessitates the use of hybrid technologies to support the requirements of data-intensive research. The chapter also describes current scientific applications that utilize a combination of technologies and discusses some future research directions.

DOI: 10.4018/978-1-4666-6178-3.ch019

INTRODUCTION

Data-intensive science emerged as a fourth paradigm after the three interrelated paradigms of science: empirical, theoretical, and computational (Hey, Tansley, & Tolle, 2009). Empirical science became the dominant form of discovery in the seventeenth century as "natural philosophers" used careful and systematic descriptions of natural phenomena gathered by direct observation to produce knowledge about the world. This was followed by theoretical science; scientists like Newton and Einstein developed general theories that have explanatory power over complex phenomena. When systems become too complex for humans to analyze without mechanical support, computational techniques such as simulation were used to assist scientists in their work.

There has been a shift in the methodology of science driven by the massive growth in data that we are now able to capture through complex new instruments. For example, the Large Hadron Collider of the European Organization for Nuclear Research is able to produce 15 petabytes (15 million gigabytes) of data annually (European Organization for Nuclear Research, 2008). The PubChem archive of National Institute of Health provides information on the biological activities of small molecules. As of August 2011, the archive contains 85 million substance records representing over 30 million chemically unique compounds (National Center for Biotechnology Information, 2011). The Securities Industry Research Center of Australia (Sirca) maintains the Australian Equities Tick History archive, which contains records of activity on the Australian Securities Exchange since 1987, including all order book entries, modifications, cancellations, and trades for supported financial instruments, time stamped with millisecond precision (Sirca, 2011). Jim Gray (2009) has described this data-intensive research as the fourth paradigm of science. This reflects an epistemic shift from computational models of science that use theoretical and mathematical models where there is a level of complexity in a problem that makes empirical observation impossible, to a form of science that has an over abundance of data that requires algorithmic process to generate meaningful results. Due to the massive amounts of available data, scientists working in data-intensive areas are reliant on information technology (IT) infrastructure and tools to extract useful information from their datasets.

This chapter is organized as follows: the background section will examine the activities and processes of data-intensive science and discuss the reasons for using scientific software. The following section describes four essential requirements of efficient scientific applications. This is followed by an in-depth review of current enabling technologies that can be used to develop applications that address the essential requirements for data-intensive science. The final sections will discuss current trends in scientific applications and propose some future research directions.

BACKGROUND

Activities performed by scientists in data-intensive fields do not normally occur in isolation, they are part of coordinated efforts to extract knowledge or meaning from data. These activities can be thought of as a "pipeline" (Hey et al., 2009) through which knowledge is produced, or more generally a *scientific process*. Scientific processes cover a wide range of activities including data acquisition, data manipulation, and the publication of analysis results. An important part of the typical scientific process is the "analysis pipeline" (Szalay,