Chapter 95
Scientific Data Management and Visualization:
A Service-Driven Integration Approach

Mariana Goranova
Technical University of Sofia, Bulgaria

ABSTRACT

One of the challenges of modern science is data exploration (eScience) that synthesizes theory, experimentation, and computation with advanced data management and statistics. The scientific community produces and consumes massive volumes of unstructured and heterogeneous data from various data sources. State-of-the-art research in "intelligent labs" explores scientific data management and visualization in distributed and heterogeneous environments. The goal of this chapter is to propose and describe a scientific data management and visualization system for scientists to perform specialized data browsing, processing, and visualization using a service-driven integration approach. In order to make scientific data more usable from the Internet, a SOA-based system that uses Web services to manage data is proposed. This chapter discusses the methodology to describe and access scientific data from various sources with different formats, and transform raw data into standard datasets that can be analyzed, processed, and visualized in an effective manner.

INTRODUCTION

Service-driven approaches have been promoted as one of the most promising trends in modern IT system design. It has been widely investigated due to the great potential of SOA (Service-oriented Architecture). SOA is based on three major technical concepts: services, interoperability, and loose coupling (Josuttis, 2007). The SOA paradigm provides the ability to locate and invoke a service across machines and organizational boundaries, both in a synchronous and an asynchronous manner. It is used in large distributed systems for the realization of business processes.

The scientific community produces and consumes massive volumes of unstructured and heterogeneous data from various data sources. Scientists need access to distributed computing and data sources and support for remote access to expensive multi-national specialized instruments. They need effective software for querying data, manipulating data, and performing data analysis and visualization. The interaction between computer science, various sciences, and engineering becomes essential for the automation of scientific data processing.

Three basic activities define the modern data science: *data capture and validation, data curation*, and *data analysis* (Hey, Tansley, & Tolle, 2009). Scientific data come in different scales and formats, from experiments and simulations. Curation fits data into the right data structures, transforming the raw data into standard datasets for scientific use. Data analysis uses databases; and modeling, analysis, and data visualization tools to mine and present knowledge for easy consumption. The massive amounts of scientific data require new processing and management tools; new paradigms to easily capture, organize, analyze, discover, interactively visualize, and publish data; and the ability to share the datasets among institutions and labs. The common strategy is to turn data into reusable services that are available as logical modules, each with a standards-based interface. This allows scientists to access and use the data more easily and improves data visibility and promotes greater reuse of data.

State-of-the-art research in "intelligent labs" explores architectures and tools for scientific data management and visualization in distributed and heterogeneous environments. Service-driven approaches are increasing used to provide functionality to scientists to easily capture, organize, analyze, discover, visualize, and publish data over the Web using services.

In this chapter, we propose a flexible, dynamic, and automated service-driven approach to access a range of tools and functionality using a service-oriented architecture. The solution is composed of specialized services, which access scientific data from different data sources with various formats, and transforms the raw data into standard data sets that can be analyzed, processed, and visualized for scientific research. The proposed data model enables easy conversion of the raw data description into a canonical form. Visualization with highly interactive abilities is an essential part of scientific exploration and analysis. The proposed approach permits Web based access allowing scientists to describe and process large quantities of their data in an easier and faster way with limited technical background in data management.

The contributions in this chapter are primarily in the nontrivial process of describing scientific observational data because of various data formats which require sophisticated processing techniques. Understanding of observational data begins by understanding the information regarding the origins, ownership, metadata, and structure layout of datasets. The development of appropriate language abstractions for science that mix compilation and interpretation; integrate code written in different languages into a unified service; and provide ways to describe the format, structure, and semantic content of data, is essential for automated data manipulations.

The organization of this chapter can be summarized as follows: In the background section, we describe the related works in scientific data management. Next, we propose an XML-based language for scientific data description and a solution architecture model based on SOA for management and visualization of scientific data. Then, we describe the implementation of a software instrument that implements the proposed architecture. The relevance of this approach is proven by applying it to real data. Finally, we highlight some future research work and conclude the chapter.

Related Content

Dynamic Research on Youth Thought, Behavior, and Growth Law Based on Deep Learning Algorithm
Qi Fu (2023). *International Journal of Data Warehousing and Mining (pp. 1-19).*
www.irma-international.org/article/dynamic-research-on-youth-thought-behavior-and-growth-law-based-on-deep-learning-algorithm/333518

HASTA: A Hierarchical-Grid Clustering Algorithm with Data Field
Shuliang Wangand Yasen Chen (2014). *International Journal of Data Warehousing and Mining (pp. 39-54).*
www.irma-international.org/article/hasta/110385

A BPMN-Based Design and Maintenance Framework for ETL Processes
Zineb El Akkaoui, Esteban Zimányi, Jose-Norberto Mazónand Juan Trujillo (2013). *International Journal of Data Warehousing and Mining (pp. 46-72).*
www.irma-international.org/article/bpmn-based-design-maintenance-framework/78375

A Study of Friendship Networks and Blogosphere
Nitin Agarwal, Huan Liuand Jianping Zhang (2009). *Handbook of Research on Text and Web Mining Technologies (pp. 646-669).*
www.irma-international.org/chapter/study-friendship-networks-blogosphere/21749

Governance in European Hospitals: Analysing Governance Practices Using Data Mining and Information Visualisation
Kristof Eeckloo, Luc Delesieand Arthur Vleugels (2009). *Social and Political Implications of Data Mining: Knowledge Management in E-Government (pp. 88-105).*
www.irma-international.org/chapter/governance-european-hospitals/29066