Chapter 3
Micro-Services:
A Service-Oriented Paradigm for Scalable, Distributed Data Management

Arcot Rajasekar
University of North Carolina at Chapel Hill, USA

Mike Wan
University of California at San Diego, USA

Reagan Moore
University of North Carolina at Chapel Hill, USA

Wayne Schroeder
University of California at San Diego, USA

ABSTRACT

Service-oriented architectures (SOA) enable orchestration of loosely-coupled and interoperable functional software units to develop and execute complex but agile applications. Data management on a distributed data grid can be viewed as a set of operations that are performed across all stages in the life-cycle of a data object. The set of such operations depends on the type of objects, based on their physical and discipline-centric characteristics. In this chapter, the authors define server-side functions, called micro-services, which are orchestrated into conditional workflows for achieving large-scale data management specific to collections of data. Micro-services communicate with each other using parameter exchange, in memory data structures, a database-based persistent information store, and a network messaging system that uses a serialization protocol for communicating with remote micro-services. The orchestration of the workflow is done by a distributed rule engine that chains and executes the workflows and maintains transactional properties through recovery micro-services. They discuss the micro-service oriented architecture, compare the micro-service approach with traditional SOA, and describe the use of micro-services for implementing policy-based data management systems.

DOI: 10.4018/978-1-61520-971-2.ch003

INTRODUCTION

Traditional data management requires the application of administrative functions to enforce management policies such as backup, retention, and disposition, and to validate assessment criteria such as authenticity, integrity, and chain of custody. The administrative functions require the management of state information about each file including the location, owner, and access controls. Service Oriented Architectures provide mechanisms to tune environments to implement specific data management policies by chaining procedures together. We explore whether a policy-based data management environment can be created that provides the extensibility of SOA while managing state information normally associated with digital libraries. We demonstrate that data analysis environments can be tightly integrated with data management environments. Indeed, for petabyte-scale collections, it is not feasible to move the entire collection to a compute server. Data analysis procedures will need to be applied at the storage location to extract the data sets of interest. In practice, it is more effective to execute low-complexity operations (that have a small number of operations compared to the size of the data in bytes) at the remote storage location. A simple example is the extraction of a subset of a file. It is faster to extract the data subset at the storage location through partial I/O commands than it is to move the entire file to a remote compute engine. Data analysis can be significantly accelerated through the execution of services at remote storage locations.

These are the driving motivations behind the integration of data processing functions into the data management infrastructure, and the execution of the functions under the control of a service oriented orchestration. We have integrated the SOA paradigm with collection management functions within the integrated Rule Oriented Data System (iRODS), and applied the technology in support of data sharing environments, data processing pipelines, data publication systems, data preservation systems, and data federation environments for long-term sustainability.

Massive Data Collections

We address the data management challenges of large-scale data systems that manage Petabytes of data and store hundreds of millions of files in a distributed environment composed from heterogeneous storage resources ranging from file systems to archives to relational databases. We automate execution of administrative functions, enforce management policies, and validate assessment criteria in order to minimize the amount of labor needed to manage massive collections. A generic solution is needed that is capable of handling discipline-centric data and catalog services for data types from high-definition video to real-time sensor data streams to simulation output. The diversity of support requirements - from small to large file sizes, from blobs to highly structured files, from static files to dynamic and active data streams, from single user systems to large-scale, distributed community-sharing networks, from free, public sites to systems with high levels of authentication and authorizations – poses a challenge that needs an intelligent and integrated operating system, executing coordinated workflows on collections of millions of digital objects stored across wide-area networks. Service-oriented architectures provide a solution for tackling this problem and catering to the distributed processing needs of such large-scale data management systems. Examples of large-scale data sharing systems at the large enterprise level can be found in business organizations as well as in the scientific/academic communities. The following provides a small list of such scientific data systems that are currently being assembled and that will exceed several Petabytes in size in production operation.

- Astronomical data: The National Virtual Observatory (US National Virtual