Environmental Modelling & Software 102 (2018) 233–240

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

# Advancing distributed data management for the HydroShare hydrologic information system

Hong Yi [a], [*], Ray Idaszak [a], Michael Stealey [a], Chris Calloway [a], Alva L. Couch [b], David G. Tarboton [c]

[a] *Renaissance Computing Institute (RENCI), University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
[b] *Tufts University, Medford, MA, USA*
[c] *Utah State University, Logan, UT, USA*

ARTICLE INFO

*Article history:*
Received 29 May 2017
Received in revised form
11 November 2017
Accepted 13 December 2017

*Keywords:*
Distributed data management
Big data
Data sharing
Hydrologic information systems
Collaborative environment
iRODS

ABSTRACT

HydroShare (https://www.hydroshare.org) is an online collaborative system to support the open sharing of hydrologic data, analytical tools, and computer models. Hydrologic data and models are often large, extending to multi-gigabyte or terabyte scale, and as a result, the scalability of centralized data management poses challenges for a system such as HydroShare. A distributed data management framework that enables distributed physical data storage and management in multiple locations thus becomes a necessity. We use the iRODS (Integrated Rule-Oriented Data System) data grid middleware as the distributed data storage and management back end in HydroShare. iRODS provides a unified virtual file system for distributed physical storages in multiple locations and enables data federation across geographically dispersed institutions around the world. In this paper, we describe the iRODS-based distributed data management approaches implemented in HydroShare to provide a practical demonstration of a production system for supporting big data in the environmental sciences.

## 1. Introduction

To enable more rapid scientific advances and discovery, it is critical to enable reproducible science and facilitate the ability for scientists to share their work and build on the work of others. Effective data discovery and reuse requires a comprehensive cyberinfrastructure that enables diverse data with different types to be annotated, discovered, accessed, visualized, analyzed, published, and serves as a platform for collaboration. HydroShare (http://www.hydroshare.org) is one example of that cyberinfrastructure; it is a web-based collaboration system for cataloging and sharing hydrologic data, models, and tools to enable more rapid advances in hydrologic understanding via collaborative data sharing, analysis, and modeling. Using HydroShare, scientists can easily discover, access, share, and collaboratively analyze hydrologic data and models, and hence accelerate hydrologic scientific discovery. Development of HydroShare was funded by the U.S. National Science Foundation (NSF) through its Software Infrastructure for Sustained Innovation program (awards 1148453 and 1148090, 2012–2017). The Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI) has assumed responsibility for operation of HydroShare as one of its water data services, and a new NSF collaborative award (1664061, 1664018, 1664119, 2017–2021) will further advance the modeling, collaboration, storage and extensibility capabilities of HydroShare. This paper reports on the HydroShare distributed data storage and management system developed using an open source data grid middleware called iRODS (Integrated Rule-Oriented Data System) (Moore (2008); Russell et al. (2016)).

HydroShare is constructed from several coupled components, including a Django front end (written in the programming language Python) that serves as a user interface, an iRODS distributed file system back end that stores files and data, a SOLR search engine (http://lucene.apache.org/solr/) that enables data discovery, and a variety of application servers that access iRODS and SOLR via a Representative State Transfer (REST) application programming interface (API). This loose coupling of application servers allows extensibility as needs of hydrologic science evolve over time (Refer to Section 2 and Fig. 1 for a more detailed description of these coupled components in HydroShare.).

[*] Corresponding author.
*E-mail address:* hongyi@renci.org (H. Yi).

https://doi.org/10.1016/j.envsoft.2017.12.008
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

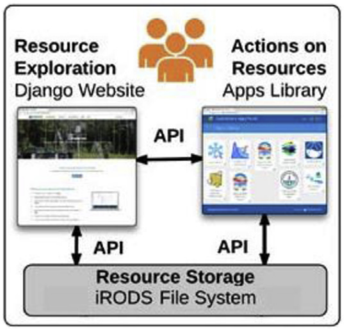

**Fig. 1.** HydroShare component architecture.

Hydrologic data and models can be large, often at multi-gigabyte or terabyte scales. To cope with large data sets, we use the iRODS data grid middleware as the distributed network data storage and management back end in HydroShare. iRODS provides a unified virtual file system for physical storage distributed across multiple locations and enables data federation across geographically dispersed institutions around the world. Using iRODS enables HydroShare to work with large files efficiently by leveraging iRODS parallel file transfer capability and addresses big data management challenges including system expansion, ability to deliver data to analysis applications efficiently, efficient data packaging, and off-site data replication for disaster recovery.

iRODS provides a distributed virtual framework for managing physical storages across multiple locations and enables storage federation across geographically dispersed institutions around the world. It is not a prepackaged solution, but rather, a middleware with pluggable architecture that supports developer-customizable policies at every point of the data management life cycle (Russell et al. (2016)), so that users are not limited by a pre-defined set of features. The flexibility and extensibility enabled by this pluggable architecture in iRODS allows us to move time-consuming data operations to iRODS, and to customize HydroShare quota management policies in iRODS. In addition, the storage federation capability in iRODS allows us to create a federated data storage system in HydroShare so that partner institutions can share the burden of storing large data sets.

In the following sections, we first provide background for our work, then describe how iRODS is used in HydroShare for managing big data in hydrological and environmental sciences. In addition, we discuss the strengths and weaknesses of our data management approach in the context of HydroShare. This is followed by discussion of the approach, use cases, and future work. Finally, we summarize our contributions and conclude the paper.

## 2. Background

An overview of the functionality envisioned for HydroShare was given in Tarboton et al. (2014) and the initial software architecture for HydroShare was described in Heard et al. (2014). Idaszak et al. (2016) provided a case study of the application of modern software engineering to HydroShare. Horsburgh et al. (2015) described the data model and content packaging scheme for diverse hydrologic datasets and models used by HydroShare to enable storage, management, sharing, publication, and annotation of the diverse types of data and models used by hydrologic scientists. These diverse types of data and models are packaged into a resource bag using the BagIt File Packaging Format (Boyko et al. (2012)) for storing resources on disk and for serializing them to zipped files for transfer. We refer to this resource packaging operation as *resource bagging* hereafter for easier reference.

A *resource* is the granular unit used for data management and access control within HydroShare. Physically, a resource is a directory in a file system that adheres to the structure of the BagIt format. HydroShare uses a resource-centric approach in which resources are objects that can be created, stored, modified, versioned, shared, annotated, discovered, accessed, published, and acted upon by web apps independent of HydroShare. As shown in Fig. 1, HydroShare's functionality and architecture can be organized into three categories: (1) iRODS-based distributed resource storage and management, (2) resource exploration, and (3) actions on resources. Each is implemented using system components that interact through APIs. The loose coupling between HydroShare and web apps enables extensibility in that anyone can develop a web app that interacts with resources stored in HydroShare. An example of a web app for visualizing spatial data in HydroShare is described in Crawley et al. (2017).

### 2.1. Related work

Vitolo et al. (2015) recently gave an overview of using web-based technologies for processing big data in the context of environmental sciences and highlighted the fact that using web and cloud-based technologies for big data analysis is increasingly acknowledged in the environmental community. There has been increased development and availability of data repositories and archival systems for the last several years, including FigShare (Hane (2013)), the NSF (OCI 0940841) supported DataNet Federation Consortium (http://datafed.org/), DataONE (Cao et al. (2016)), SEAD (Myers et al. (2015)), and CyVerse (Merchant et al. (2016); Oliver et al. (2013)). These systems all have a similar basic structure: a data store, the ability to “publish” data for public consumption, and ability to specify metadata for published data so that it can be discovered by other researchers involved in similar tasks.

HydroShare enables interoperability with other digital repositories via a standards-based approach to data storage and metadata. HydroShare's data model (an adaptation of the Open Archive Initiative's Object Reuse and Exchange (OAI-ORE) standard Lagoze et al. (2008)), metadata structure (Dublin Core with extensions DCMI (2012)), and packaging scheme (the BagIt hierarchical file packaging format) all use well-known standards within the digital preservation and archival community. HydroShare's structured and comprehensive resource types and metadata descriptions not only facilitate better interpretation by users, but also enable users to write independent applications designed to operate on this structured content.

Unlike many other data publication systems, HydroShare enforces a strict compliance to metadata standards at all times, and strict synchronization between metadata and the presence of specific resource files. The reason for this requirement is that HydroShare allows authorized users to treat the resource as if published at any time, and download, use, and then re-upload the resource as part of the data life cycle, even before the resource is formally published. By contrast, many publication systems prohibit download until metadata is synchronized with object contents. That mode of operation would interfere with a core goal of HydroShare: to allow resource sharing *before* (or even without) formal publication.

HydroShare can also interoperate with other data repositories via its iRODS-based distributed data storage and management system. iRODS is used by large scientific research projects across the country and around the globe for managing petabytes of data in hundreds of millions of files on distributed storage resources. For example, Hedges et al. (2009) presented a rule-based data grid approach using iRODS for automatic data curation and