Service Oriented Architecture for Harvesting Distributed Data Repositories

Helen Conover, Sara Graves, Ken Keiser, Lamar Hawkins
University of Alabama in Huntsville
Huntsville, Alabama 35899

Abstract - Through the use of a service oriented architecture, the REASoN DISCOVER project is incorporating the ability to harvest inventory metadata from distributed repositories to be updated in the systems data catalog for GHRC Data Pool interoperability across these distributed repositories. The harvesting workflow incorporates the deployment of a light-weight harvesting agent at each of the distributed repositories that scans the repository at scheduled intervals for all data granules, producing an inventory metadata index. The Harvester agent then uses a SOAP web service to communicate with a centralized "Combine" component that a new index is available to be processed into the catalog's inventory and the web-accessible location of the index. Once triggered, the Combine retrieves the inventory index from the remote location and processes the metadata into the centralized catalog as appropriate, adding new inventory items, updating existing items, and removing references to inventory no longer available. This Combine processing keeps the GHRC Data Pool's catalog up-to-date with the contents of the remote repositories.

Initial versions of the Harvester have been deployed and tested at both the GHRC and Remote Sensing Systems data repositories. The goal is to keep the distributed Harvester agent component as lightweight and easily deployable as possible to avoid unnecessary installation complexity at the existing and future repositories. Currently the Harvester is written in Perl, and as such is portable to most systems. Since the GHRC Data Pool is focusing on "online" repositories, the availability of web-accessible locations for the Harvest index should be a reasonable requirement. The use of SOAP for the service protocols will allow the catalog services to be incorporated with other systems that are employing service architectures. An overview of the harvesting technology will be presented.

I. Introduction

The Global Hydrology Resource Center (GHRC) has been investigating the use of web services for managing distributed repositories of scientific data for the REASoN DISCOVER (Distributed Information Services for Climate and Ocean Products and Visualizations for Earth Research) Project. The GHRC is a collaboration between the Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville and the Global Hydrology and Climate Center (GHCC) at the National Space Science Technology Center in Huntsville, Alabama. DISCOVER is a collaboration of Remote Sensing Systems (RSS), the Global Hydrology & Climate Center (GHCC) and the University of Alabama in Huntsville (UAH). The primary objective of the DISCOVER Project is to provide highly accurate, long-term ocean and climate products suitable for the most demanding Earth research applications via easy-to-use display and data access tools. A key element of DISCOVER is the merging of data from multiple sensors on multiple platforms into geophysical data sets consistent in both time and space. The information technology focus of DISCOVER is on providing services for online data access, ordering and visualization of the project's data products and information.

The GHRC Data Pool [1,2,3,4] provides a data search-and-order interface for data in distributed repositories maintained by a single management infrastructure. The next generation of this data pool is integrating data from repositories maintained by different institutions so it has been necessary to develop tools and functionality capable of consistently collecting and managing metadata across the distributed online resources. A self-imposed constraint is the desire to have minimal impact on the repository's workflow.

II. Approach

The foundation for an information management system is complete metadata describing a repository's contents. The DISCOVER project is providing automated metadata generation for distributed data repositories containing heterogeneous online data products. The solution being implemented for this project involves deployable on-site utilities (agents) for metadata generation and web services for distributed metadata collection and catalog interactions. The main components of this service oriented architecture include the distributed Harvester for gathering metadata at the repositories' locations, the Combine that handles the comparison of harvested metadata with the catalog contents and finally a suite of catalog services for metadata updates, synchronization and maintenance.

A. Harvesting Metadata

In the GHRC local repositories, as with other archives, metadata is typically collected as part of the ingest processing where inventory information is updated or removed as changes occur. The GHRC Data Pool is integrating metadata from distributed repositories which

have their own data management workflow and the goal is to minimize impact on those remote systems and processes. Towards that end, we have adopted a web crawler approach to indexing the distributed repositories, with the permission of those facilities. The Harvester component, or agent, is deployed on-site at participating repositories with minimal site-specific configuration to target the harvesting of metadata for specific data collections at set time intervals (see Figure 1).

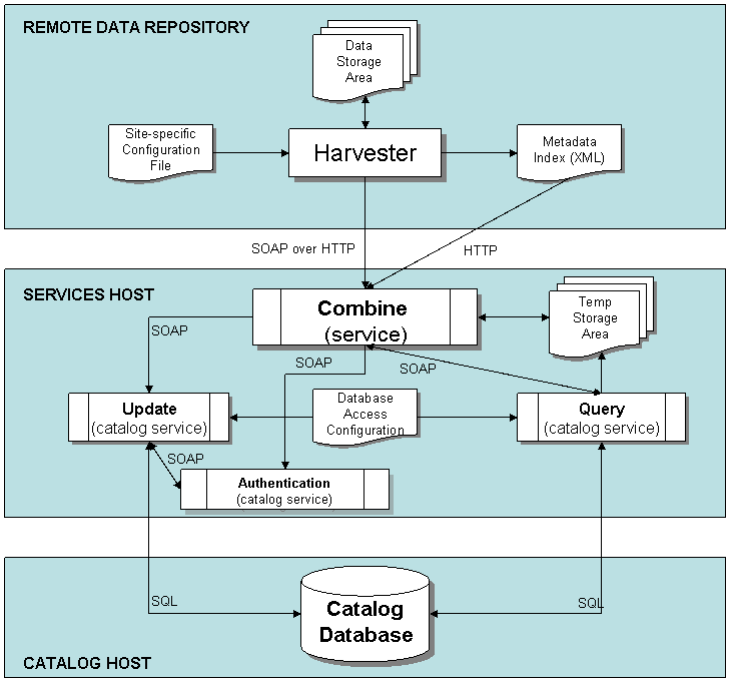

Figure 1: Harvester-Combine Functional Diagram

To minimize network traffic and possible security concerns, the Harvester generates an index of the repositories' inventory and associated metadata in a local file created in a web-accessible (HTTP) location. Only one distributed service (SOAP) is called from the Harvester to tell the remotely located Combine that a new index is available, and provide the URL of the new file. This approach also allows the Combine to efficiently handle input from multiple Harvesters by brokering resources at the central location and decide when best to process the new inventory information.

The collected Metadata Index is organized based on an XML schema [5] defined to generically handle basic inventory information for any data collection.

B. Combining Metadata

The Combine handles integrating the remotely harvested metadata indexes with a central catalog. The catalog is synchronized to represent the indexed contents of the distributed repositories, allowing the catalog-driven Data Pool search-and-order interface to provide current inventory information to the users. As depicted in Figure 1, the Combine retrieves the metadata index from the remote repository after being notified that a new file is available. The Combine figures out from the index what data collections are involved, allowing it to *query* existing inventories and then perform a comparison between the new index and the catalog's current inventory. The differences are then noted and the necessary *updates* (additions/modifications/deletions) are performed on the catalog. Since the updates are "write" actions on the database, the Combine must *authenticate* the repository's catalog access prior to processing the new index. The query, update and authenticate services are part of a suite of catalog services described in section 'C' below. The Combine is a distributable component [7] that could be implemented locally or remotely from the catalog, and there could be multiple Combines providing load distribution of resources if necessary.

C. Servicing the Metadata Catalog

A suite of catalog services, including the query, update and authenticate services depicted in Figure 1, are employed by the Combine to facilitate communication and interactions with the data catalog. SOAP interfaces have been deployed for these services to make them readily available to other applications and users as necessary. The service interface definitions are publicly available [6], but user authentication is required for any of the services that require "write" access to the catalog.

Summary

The Harvester-Combine approach effectively utilizes a service oriented architecture to facilitate the remote harvesting of inventory information and the resulting metadata catalog updates. This distributed service approach insulates the various components from platform and language dependencies, with the only language-dependent component being the deployable Harvester. The Harvester was written in Perl, utilizing platform-independent file system access modules so it has proven to be portable on all tested systems, to date. We believe the distributed service architecture will also prove to be scalable as the various components can be deployed at multiple locations to maximize resource utilization if load balancing should become an issue.

While this system is currently in a beta-testing phase, test results have been promising and we expect to have a production version deployed in the near future.

Acknowledgment

The authors would like to acknowledge the other DISCOVER partners at the National Space Science and Technology Center (NSSTC) and at Remote Sensing Systems, who have collaborated on the testing and deployment of these and other information technology components and processes.

References

[1] "Distributed Technologies in a Data Pool", Ken Keiser, Helen Conover, Sara J. Graves, Yubin He, Kathryn Regner, Matt Smith, American Geophysical Union 2004 Fall Meeting, San Francisco, CA, Dec. 13, 2004

[2] "Distributed Services Technology for Earth Science Data Processing", Ken Keiser, Rahul Ramachandran, John Rushing, Helen Conover, Sara J. Graves, American Meteorological Society's (AMS) 19th International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Long Beach, CA, Feb. 9, 2003

[3] "DISCOVERing Online Data and Services", Sara J. Graves, Helen Conover, Ken Keiser, NASA Earth Science Technology Conference, Palto Alto, CA, Jun. 22, 2004

[4] URL for DISCOVER's GHRC Data Pool:
http://datapool.nsstc.nasa.gov

[5] URL for Harvester Index Schema:
http://ws.itsc.uah.edu/services/combine/harvest.xsd

[6] URL for Catalog Service definitions:
http://ws.itsc.uah.edu/services/catalog/ghrc/

[7] URL for Combine Service definitions:
http://ws.itsc.uah.edu/services/combine/