The Earth System Grid Federation - Management of Distributed Data

The Earth System Grid Federation (ESGF) is a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data [1].
ESGF has addressed the problem of making large-scale data accessible in a distributed fashion when such data originates at geographically disparate sources worldwide, eg. computing centers that run climate models or sites that curate collections of observation. Instead of having to move data (upwards of 100s of Terabytes) to a centralized location, each participating institution hosts the data, but indexing services are handled remotely by select index sites. ESGF provides a software stack of services to be installed at the participating site and allow for some degree of customization. We present an overview of the system, how several of the many challenges have been meet and some thoughts on the future.

There have been challenges in the deployment of a system that incorporates distributed data, federated identity services, and replicated, distributed indexing. Some such solutions have included our "truststore" (collection of Certificate Authority roots) distribution and replicated Solr shard creation.

A particular challenge addressed for ESGF at LLNL has been the automated process of replicated data, needed to enhance the availability of such data for the world wide community, as LLNL serves at the leading center with the largest archive, expected to grow to 11 PB and beyond. Data must scanned for ingest, versions are managed in a database, added to a catalog to provide several data services and indexed remotely. The tools to accomplish these tasks were originally designed for manual operation, We have developed a workflow to automatically move data through these stages as triggered by the detection of data recently moved through use of specialized data replication platform (Synda) that interfaces with ESGF, but comes short of handling the ingest for making our replica data available through the search portal.

While ESGF has remained successful in distributing of data, and at present is meeting the goal of an even larger data volume that could very well stress our current system to limits, ESGF has reached a point in its time where we are considering significant changes. We have (1) observed that many technologies chosen have evolved, (2) seen requirements shift, and (3) learned about many aspects of the system for where we can improve. The community has focused recently on this reevaluation with some short-term and much larger long-term goals for a refresh to the architecture and constituent components of ESGF. An example of a recently adopted feature base on a technological development made since the initial adoption of ESGF has been the Ansible playbook distribution of software stack. Several other

technologies explored for ESGF, and just in their infancy as far as production deployments go have been Docker containers and public cloud deployments. One additional example of a change to be considered that could adopt newer technologies is the potential for Cloud-replicated, distributed services, with logical centralization, as opposed to the current model of federated, distributed services.

[1] https://esgf.llnl.gov/