THE EARTH SYSTEM GRID FEDERATION: INFRASTRUCTURE FOR CLIMATE RESEARCH DATA MANAGEMENT AND DISSEMINATION

The Earth System Grid Federation (ESGF) peer-to-peer enterprise system is an interagency and international effort led by the U.S. Department of Energy, and co-funded by National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, National Science Foundation, as well as international laboratories including the Max Planck Institute for Meteorology, the German Climate Computing Centre, the Australian National University/National Computational Infrastructure, and the British Atmospheric Data Center.
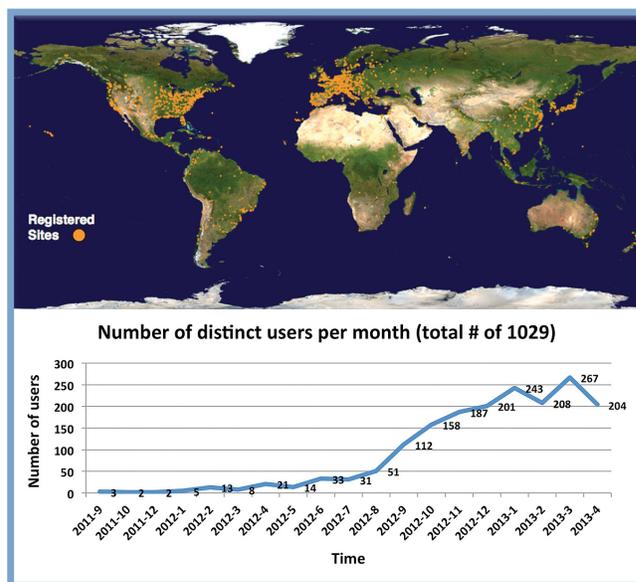
As climate data sets continue to expand in size and scope, the necessity for performing data analysis where the data is co-located (i.e., server-side analysis) is becoming increasingly apparent. Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) is undergoing greater modifications to allow ESGF's infrastructure to access its capabilities on the back-end as well as from the front-end. These modifications allow users to not only access the ESGF petabyte archives from their laptops, but also to perform analysis and data reduction before moving the data to their site for further UV-CDAT manipulations. More importantly, the necessary remote operations will be routinely performed, thus freeing researchers to concentrate on scientific diagnosis rather than on the mundane chores of data movement and manipulation.

SUPPORTING COLLABORATION

As global, broad-based climate change projections have become more useful, effectively managing the accompanying volumes of data represents a major challenge for the computational scientists who support the projections. For climate change science, this "big data" challenge is being met with ESGF, an international collaboration led by Lawrence Livermore National Laboratory. ESGF combines grid-based computing with a distributed architecture, keeping participating members sovereign while simultaneously linking them together. To achieve that, ESGF developers created a unique system of nodes that requires very little explicit coordination while still providing a robust "data space" for storage and computation.

ADVANCED TOOLS

Understanding and predicting climate change and extreme weather events requires advanced tools to securely store, manage, access, analyze, visualize, and process enormous and distributed data sets. Designed and maintained through a successful international collaboration of dozens of research institutions, ESGF now powers
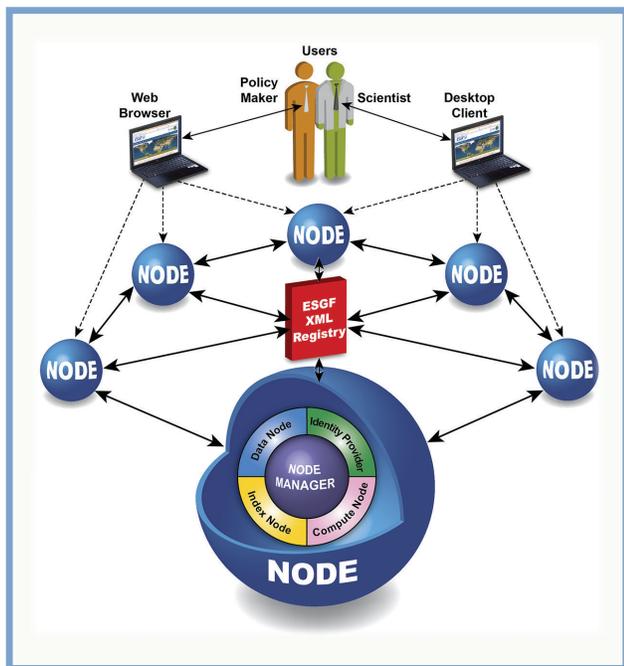

Number of distinct users per month (total # of 1029)

Currently, 25,000 users (researchers and non-researchers) from 2,700 sites on six continents are sharing data through ESGF.

most global climate change research, notably assessments by the International Panel on Climate Change. The newest iteration of ESGF offers an immense, computerized climate database that standardizes and organizes observational and simulation data from 21 countries, allowing scientists to compare models against actual observations. A rich set of climate analysis tools is available to help manipulate the data.

ESGF allows teams to work in highly distributed research environments, using unique scientific instruments, exascale-class computers, and extreme amounts of data. Users can access ESGF data using Web browsers, scripts, and client applications. A key to ESGF's success is its ability to effectively produce, validate, and analyze research results collaboratively. For example, new results generated by one team member are immediately accessible to the rest of the team, who can annotate, comment on, and otherwise interact with those results.

The ESGF architecture is based on a dynamic system of nodes—independently administered yet united by common protocols and interfaces—that interact on an equal basis and offer a broad range of user and data services. Data are published, stored, and served from dozens of nodes around the globe, yet they are searchable and accessible as if they were

*ESGF is composed of four interacting components: **data node** (publishes and serves data), **index node** (harvests metadata and enables data discovery), **identity provider** (registers, authenticates, and authorizes users), and **compute node** (contains application servers for data reduction, analyzing and visualizing data).*

The Web front-end directly accesses the ESGF archive. Integrated with ESGF, analysis and visualization tools are collocated where the data resides to optimally derive remote data products.

stored in a single global archive. Metadata shared among projects help fully integrate the repository of data and components for usability and interoperability. ESGF also promotes standard conventions for data transformation, quality control, and data validation across processes and projects.

ACCOMPLISHMENTS

ESGF is designed to remain robust even as data volumes continue to grow exponentially. More than 2 petabytes of data have been downloaded to the climate community through ESGF, making it one of the most complex, successful big data systems in existence.

FUTURE DIRECTIONS

ESGF's architecture can easily be leveraged for accessing data from other scientific domains, such as astrophysics, biology, chemistry, combustion, energy, fusion, material science, and nuclear energy. ESGF is now in the early stages of being adapted for use in two domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st century). Satisfying the need for substantial investments in software and technologies across many data-driven domains is paramount as future computing platforms and archives expand and reach extraordinary speeds and capacity.

CONTACTS

Renu Joseph, Ph.D.
DOE Program Manager
Regional and Global Climate Modeling
renu.joseph@science.doe.gov

Dean N. Williams
Principal Investigator
Lawrence Livermore National Laboratory
williams13@llnl.gov

Project Website
http://pcmdi9.llnl.gov/esgf-web-fe/