Lawrence Livermore National Laboratory

ESGF
Earth System Grid Federation

ESGF: A community resource for planet-scale data distribution and analysis

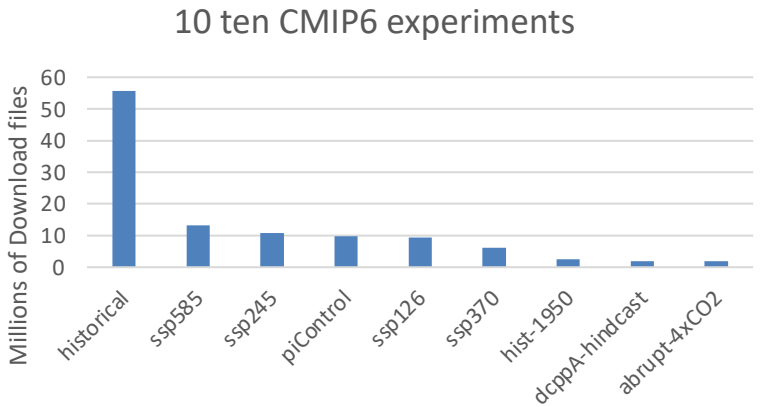
Sasha Ames

Earth System Grid Federation

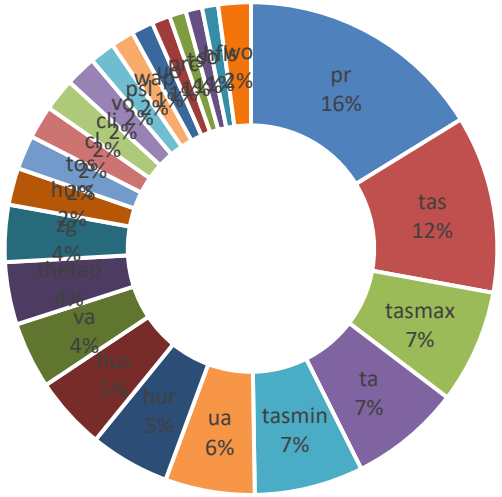
RGMA October 2020

ESGF has led data archiving for the Coupled Model Intercomparison Project (CMIP) since its conception

- Federated system facilitates data availability
  - 45 data nodes
  - 12 PB distinct data and 9 GB replica
- Publication workflow developed for modeling centers
  - In coordination with the WGCM (WIP and CMIP panel)
  - Has a detailed step by step straightforward process with available tools.
  - Allows for project updates to accommodate changes.
- Data is searchable and organized
  - Based on CMIP6 data record syntax
  - Easy to use search API
- Convenient command line download
  - wget scripts from frontend (updated API forthcoming)
  - Sproket download client
- Globus integration
- System enables a maintainable replica archive
  - Facilitates automation
  - Consistency checking improves timeliness of updates
  - LLNL makes replicas available 48hrs best case, 1-2 weeks on average after workflow optimized



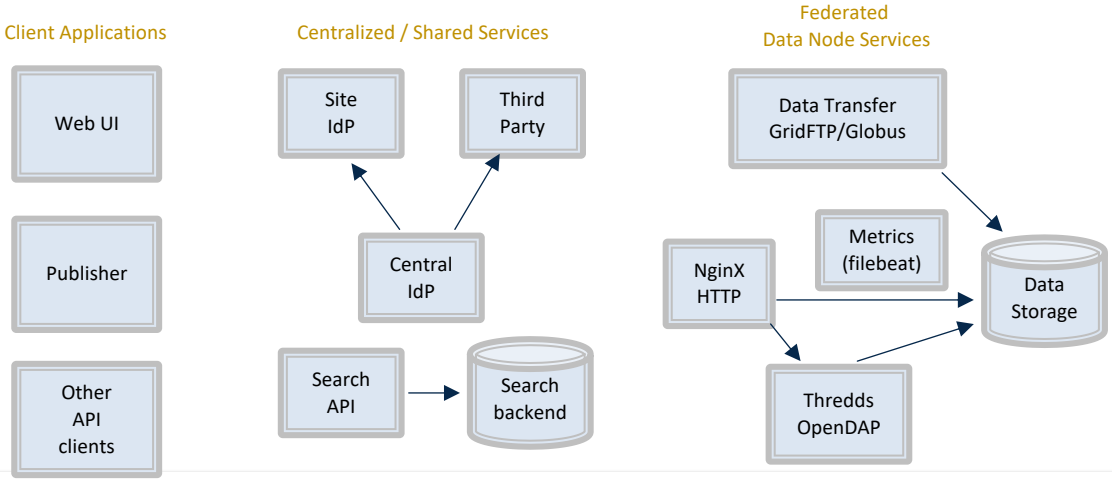
Top 24 downloaded variables (70% of total)



Our project is working to address community data management needs in enhancing the platform

- Goals and Requirements:
  - Search needs to be efficient, consistent and expressive
  - Data nodes must be reliable and scalable
  - Easy to use publishing
  - Interoperate with ES-DOC and community metadata services
  - CMIP6 data volume has pushed current system to its limits
- LLNL co-organized webinars and meetings
  - Understand user, system requirements
  - Propose solutions
  - Draft new architecture
- User facing issues and features:
  - Survey completed and themes identified
- Compatibility with cloud service providers:
  - Higher availability, support regional failover
- New publisher for emerging architecture
  - Improved modularity and extensible
  - Uses autocurator for netCDF-CF

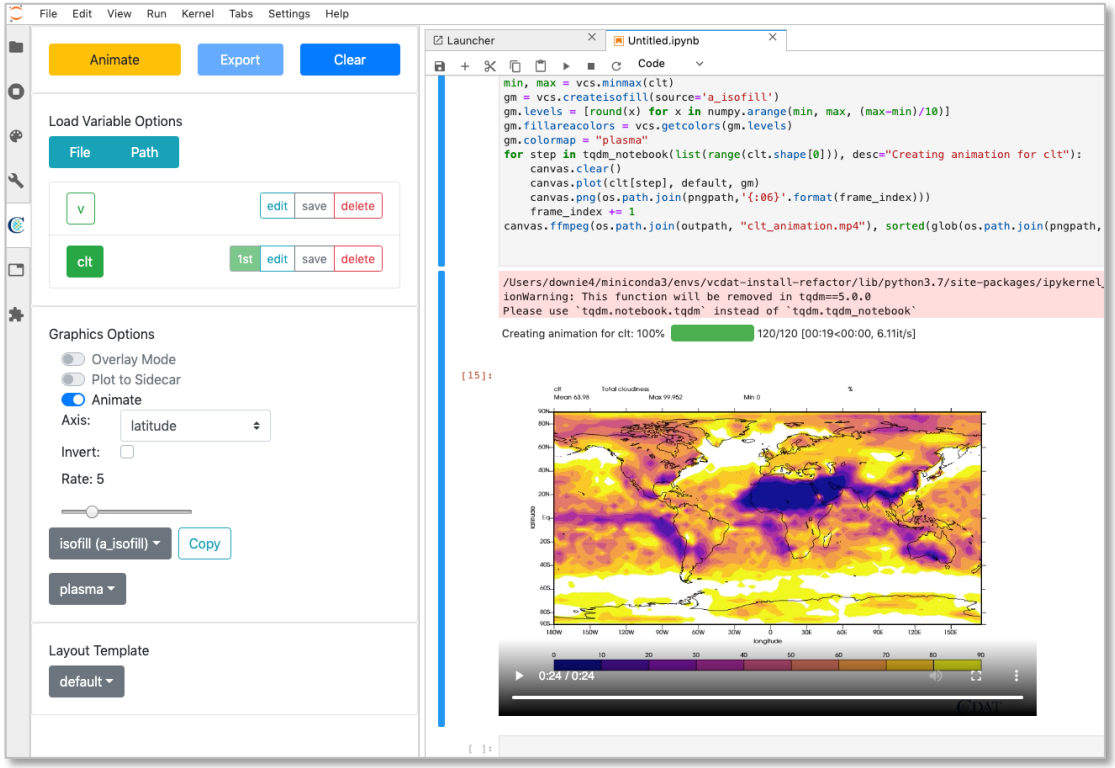
- Centralized Search and Index Services
  - Address replica record consistency concerns
  - Integrate server-side QA to benefit
  - Improved consistency
- Migration of Logins (Identity Provider/IdP) to OpenID Connect
  - Supports Third-Party logins
  - Streamlines application integration
- Simplified Data Service Deployment Mechanism
  - Docker / Kubernetes
  - Ease on operational teams
  - Improved resiliency of data services
- Ongoing Challenge: maintenance of federated data nodes
  - Comprehensive approach needed to keep sites operational and compliant



Developed VCDAT for Client-side Analysis and Visualization

VCDAT 2: JupyterLab Extension

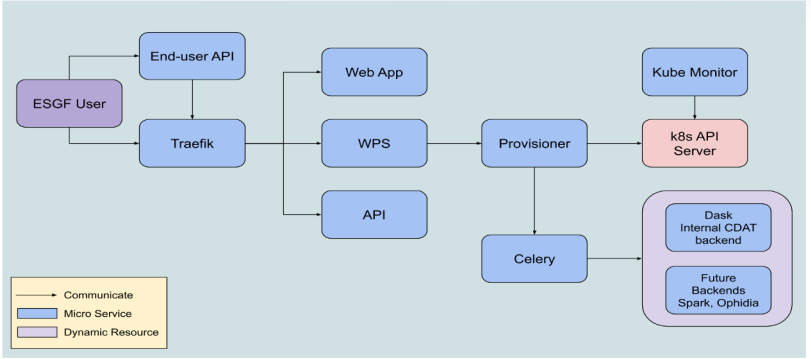
- Combines JupyterLab's feature rich interface with CDAT functionality
- Seamlessly work with the U.I. or Python code in Jupyter notebook
- Offers users installation and maintenance free experience through JupyterHub
- User survey shows users want to work with JupyterLab



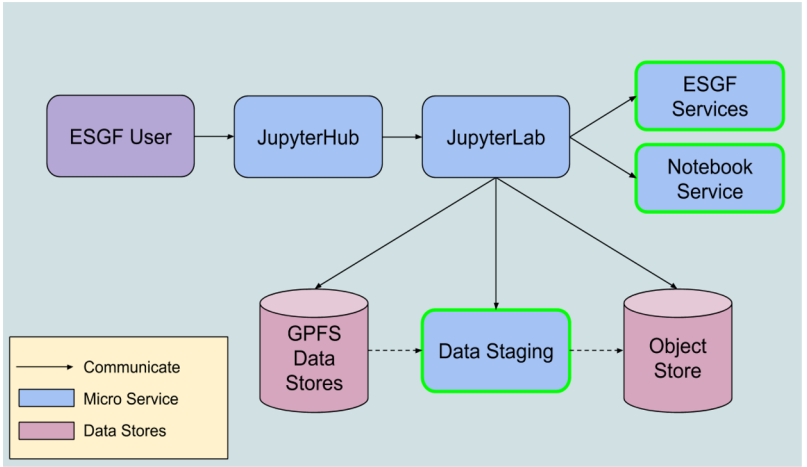
VCDAT 2 was developed to provide a more versatile and effective means of interacting with CDAT's client-side analysis and visualization tools.

ESGF Compute Node provides a scalable analysis platform available at LLNL on the Nimbus cluster

- Need for Analysis Services
  - ESGF data reached 12 PB and the average variable size ranges from 1.15 GB for input4MIPS to 3.57 GB for E3SM
  - Multiple variables per file and latitude, longitude, time coordinates
  - Data access is a big challenge because of the data sizes
- Approach and benefits
  - Enable data reduction where data resides
  - Reducing to a single dimension can result > 90% data size reduction
  - Large savings on data download and will speed the data analysis process for scientists
- Significance and impact
  - Reduce time to download relevant data for analysis and save storage space on the client side
  - Integrate with search, publishing, and errata ESGF services
  - Provide data staging service and support larger computer cluster and/or Cloud services
  - Shorten time for scientific discovery



Current architecture of the compute node to reduce data with server-side compute



JupyterLab deployment, future work to integrate ESGF ecosystem and community services

Useful links

- ESGF Software overview: https://esgf.llnl.gov/software.html
- ESGF Support Documentation:
  - https://esgf.github.io/esgf-user-support/
- ESGF Future Architecture:
  - https://esgf.llnl.gov/esgf-media/pdf/ESGF_Future_Architecture_Report.pdf
- CDAT: https://cdat.llnl.gov
- VCDAT2: https://github.com/CDAT/jupyter-vcdat/wiki
- Compute API: https://github.com/ESGF/esgf-compute-api

- Thanks!

ESGF
Earth System Grid Federation

Lawrence Livermore
National Laboratory