ESGF2-US Data Platform Modernization

Sasha Ames, Rachana Ananthaksrishan, Lee Liming, Max Grover, Jitu Kumar, Steve Turoscy, Lukasz Lacinski, Forrest Hoffman, Ian Foster and the ESGF2-US Team (ORNL, ANL, LLNL, UChicago/Globus)

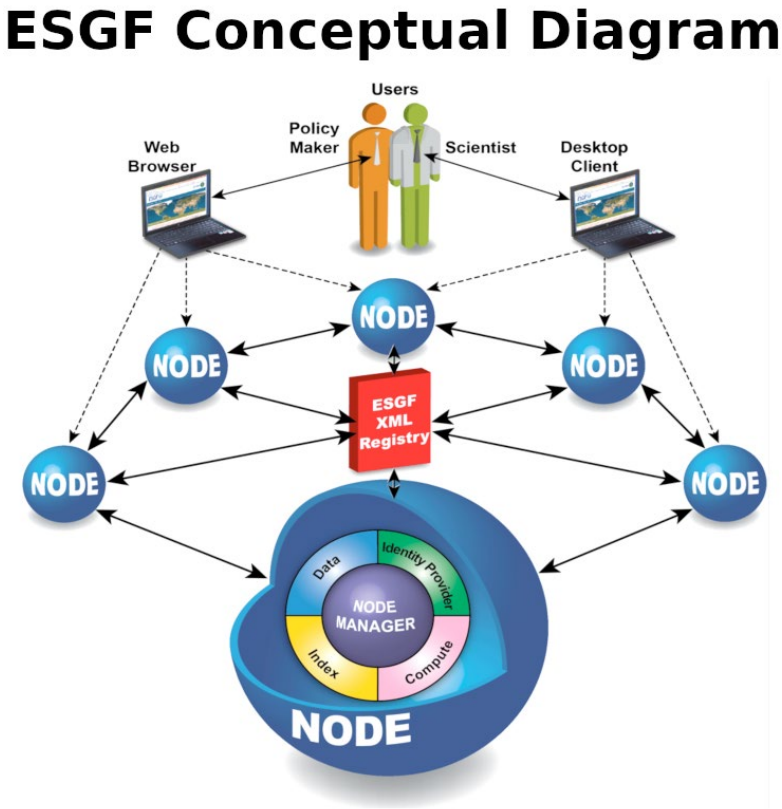
DOE Data Days - October 26, 2023

U.S. DEPARTMENT OF ENERGY | Office of Science

What is ESGF?

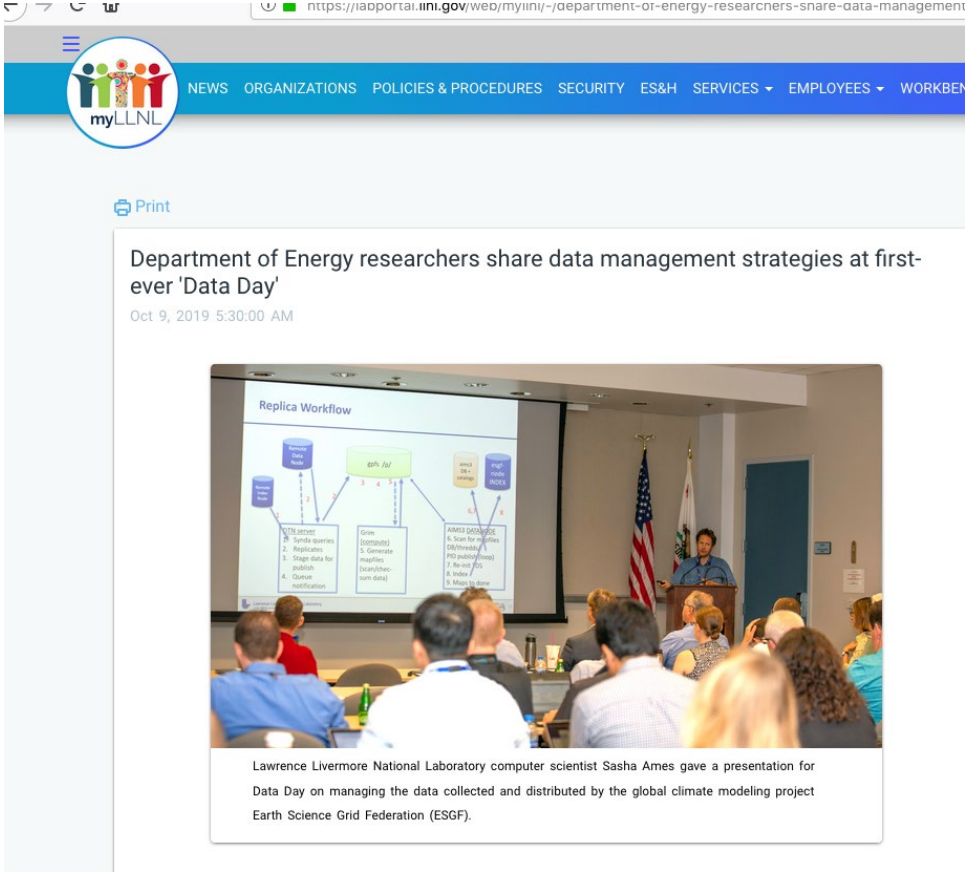
A: The Earth System Grid Federation (ESGF) is a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. ESGF's primary goal is to facilitate advancements in Earth System Science.

Traditionally, the Federation has operated a collection of data services distributed at many partner sites.



State of ESGF c. 2019

- Most services are “federated”
  - Data, Index, Identity
- Things worked well but failures were notable
  - Solr indexing reached limits
- Focus on recent trends in related technologies
  - Cloud-hosted platforms for search
  - Container orchestration
  - User computing via notebooks
  - Web application advances with SPA-style, eg. React
  - Beginnings of Analysis-ready, cloud optimized data formats, eg. Zarr.



ESGF2-US Project Overview

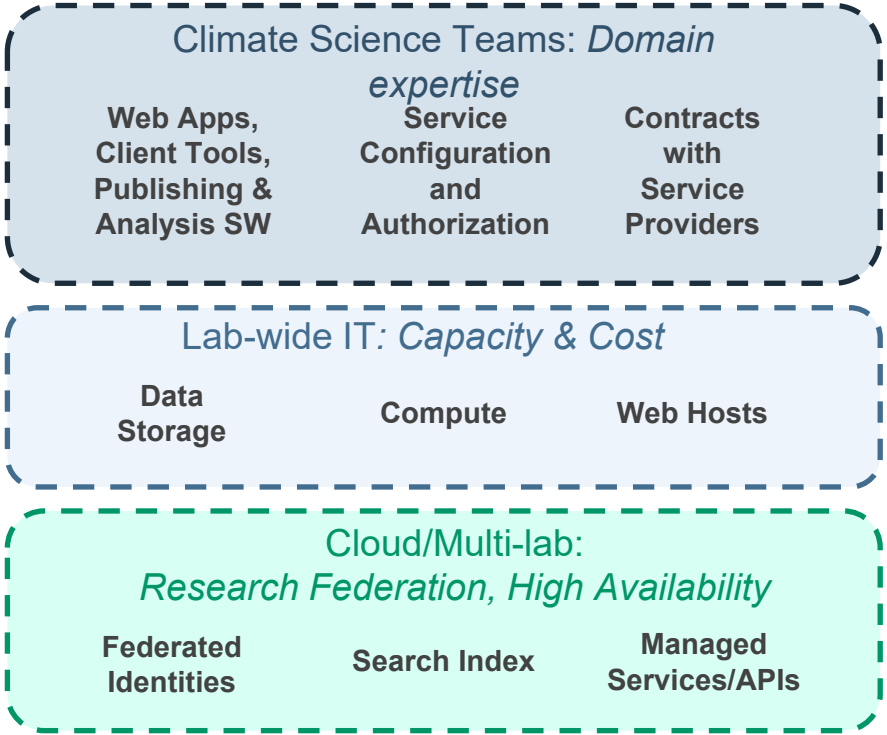
- Project began under ORNL leadership with ANL, LLNL Summer 2022
- Currently funded by US-DOE into CY2025
- Inherited from prior work led by LLNL
- Dual technical approach
  - Leverage Globus services for platform development
  - Deploy services for user data integration to containerized environment

PIs:
Forrest Hoffman (ORNL)
Sasha Ames (LLNL)
Ian Foster (Argonne)

U.S. DEPARTMENT OF ENERGY | Office of Science | OAK RIDGE National Laboratory | Argonne NATIONAL LABORATORY | Lawrence Livermore National Laboratory

ESGF 2 Trending Directions for US-DOE ESGF Data Nodes

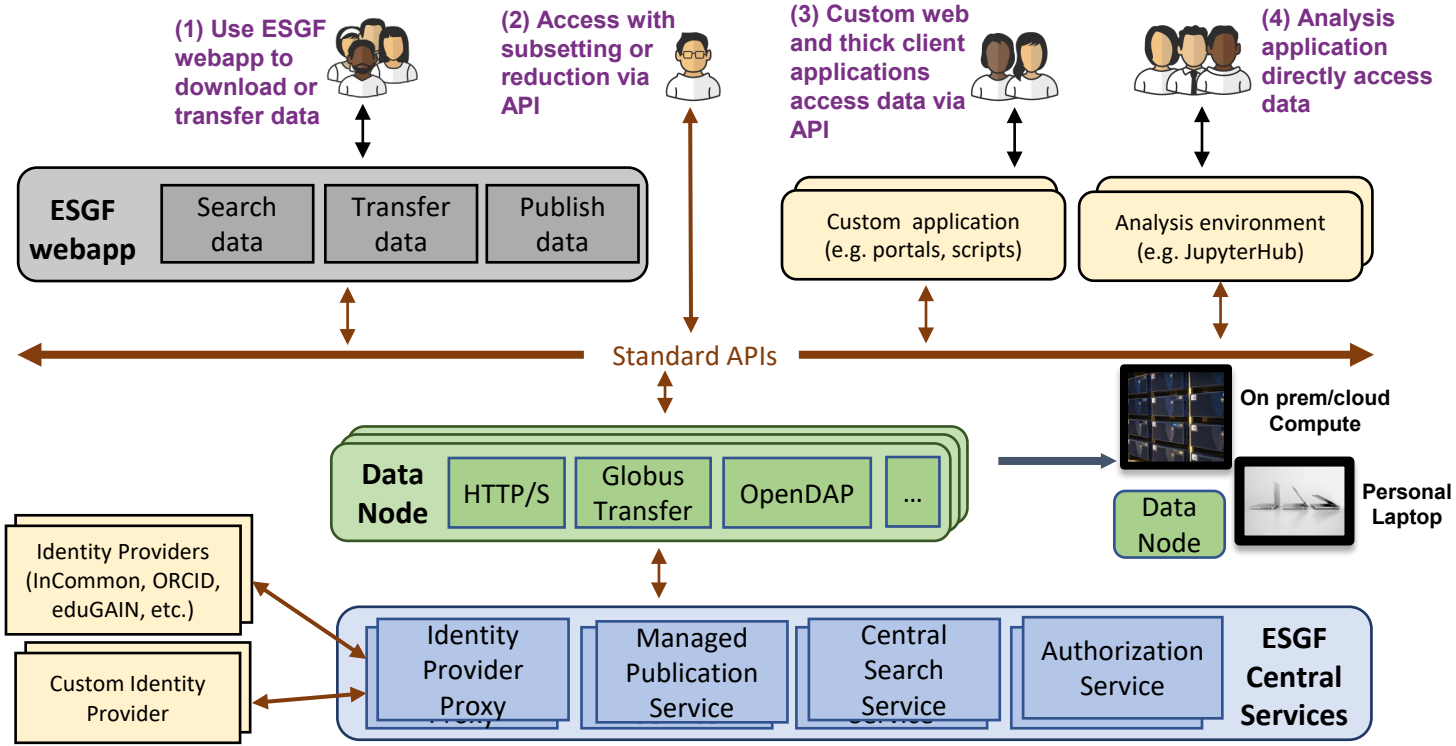
- Significantly ***increased capacity*** for ESGF in both **data storage** and **compute**
- Significantly ***more reliable ESGF APIs*** (e.g., **data access**, **discovery, authentication**)
- ***Common, agreed-upon interfaces*** across the Federation
- Trending toward ***fewer methods for direct fabric access*** (but those that remain will be robust and scalable)
- Trending toward ***more hosted applications and workflows*** and less emphasis on downloadable software



ESGF2 Data Access

- Data storage provided on ***lab-wide research systems*** operated by **lab-wide IT teams**
  - Superior **capacity and cost**, does not require ESGF personnel to operate
  - Globus Connect Server enables research IT staff to provision collections for specific scientific teams (like ESGF) with distinct access policies and team-specific configurations (including administration)
    - ESGF personnel (data managers) configure directory-level access permissions
    - Lab-wide Globus data services already support 1,000s of researchers at DOE labs, 400k+ researchers globally
    - Replicated 7 PB ESGF data from LLNL to Argonne and Oak Ridge; continuing with daily updates (CMIP6)
- ESGF collection enables **both public & restricted access**
  - **Direct HTTP/S** (download, upload, byte range access)
  - **Globus Transfer API** - highly-available bulk transfer to/from other Globus collections (35k+ globally)
  - For restricted access (e.g., publishing), Globus Auth offers superior **research federation support** via InCommon, eduGAIN, and other OpenID Connect (OIDC) integrations
  - ESGF MetaGrid web app provides UI/UX for both browser download and bulk data transfer

ESGF 2 Data Access



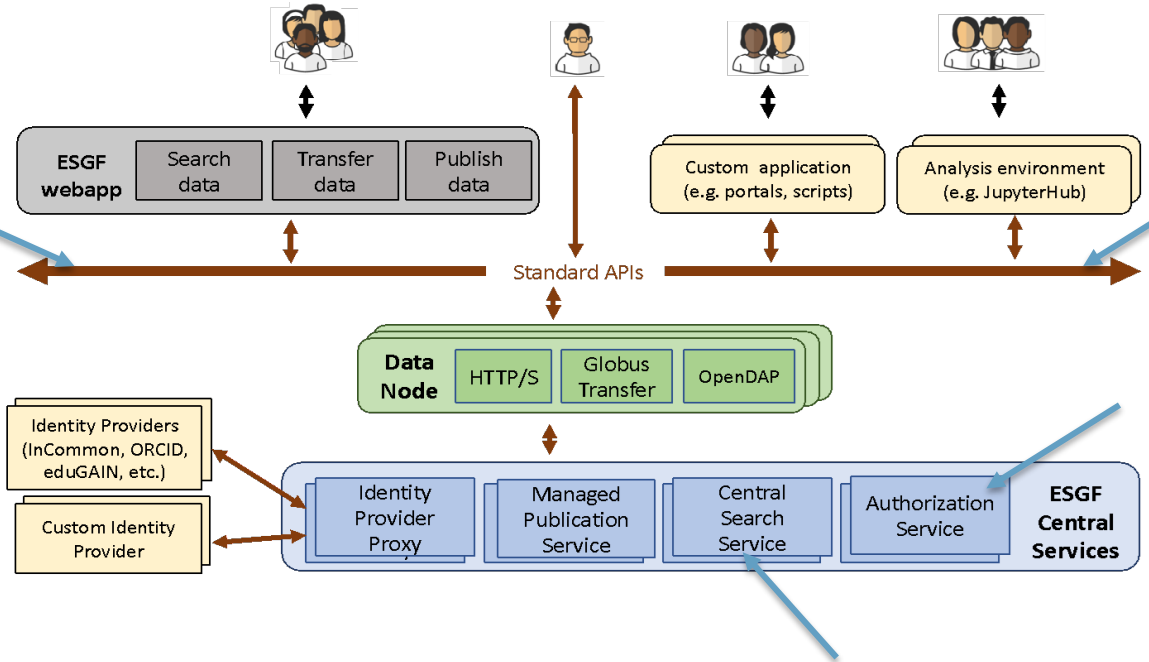
ESGF2 Data Discovery

- Aiming to operate **a single U.S. index**
  - Everything published at US/DOE data nodes (Livermore, Oak Ridge, and Argonne)
  - Ingest additional metadata for assets across the federation
  - Data compatibility – continue using ESGF metadata model/schemas
  - Interface compatibility – ESGF-agreed search interface (STAC is being investigated)
- Commercially hosted AWS ElasticSearch
  - High availability, automatic patches for vulnerabilities and bugs
  - Globus Search API adds **research federation IAM** (institutional identities, item-level visibility permissions) to ElasticSearch
    - Used by multiple research teams, operated by professional IT personnel, supported by 100s of campus/lab subscriptions
  - No US-DOE ESGF operations personnel required for index service

ESGF 2 Using Globus Search in ESGF

APIs are used to ingest metadata into the index.

Supports asynchronous bulk ingest



Discovery and Query API for faceted search.

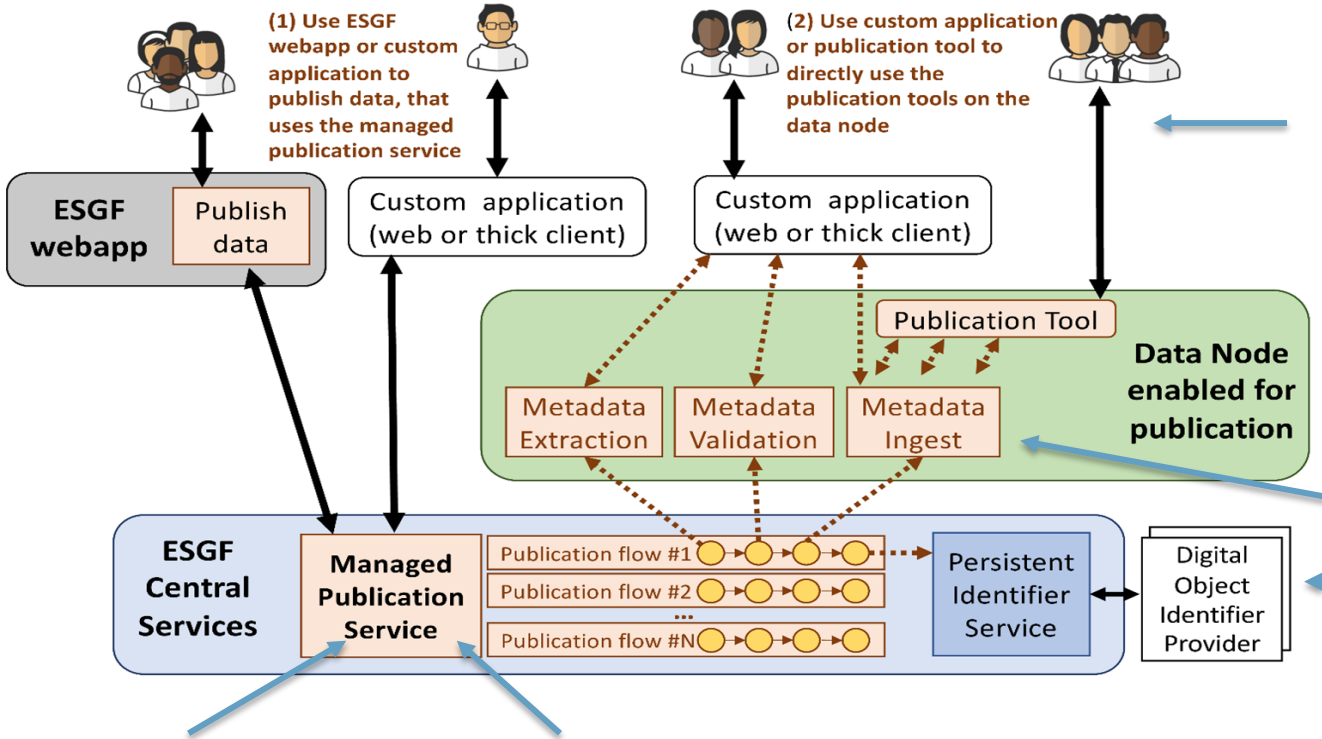
Visibility of the metadata elements can be configured on the index

An index is created and managed in Globus Search service

ESGF2 Publication and Data Processing

- Experiences
  - Individuals with publisher privilege can—using web browsers—provide inputs and publish data without downloading/installing publisher software.
  - Climate scientists can—using web browsers—perform well-known data processing workflows without transferring datasets or installing software.
- Lab-wide compute platforms
  - Superior capacity and cost
  - Complex publication & data processing orchestration managed by Globus Flows
    - Research federation IAM (via eduGAIN/InCommon)
    - Highly-available API/SDK/web app interfaces
    - Define ESGF publishing & data processing workflows that include server-side processing on lab-wide compute platforms, eg. bulk regridding
    - Authorized individuals can run workflows via web app, SDK, CLI
  - Workflow steps are software/applications and can be accessed via other clients
  - We'll offer publishing flows for well-know dataset types (e.g., CMIP6/7, E3SM, etc.)
  - Publishing and processing code remains open source, and we'll continue to support local installation & configuration for new dataset types and local publishing

ESGF 2 Data publication service architecture



Each publication submit is a run of the flow, and managed in the Flows service

Actions in the flow invoke some services hosted on the ESGF node, and other from external providers

Create a flow, per project, with the steps needed for publication, and set policy

Publication runs can be monitored and managed in the Flows service

ESGF2 US Model Data Integration and User Computing

ESGF2-US User Communities

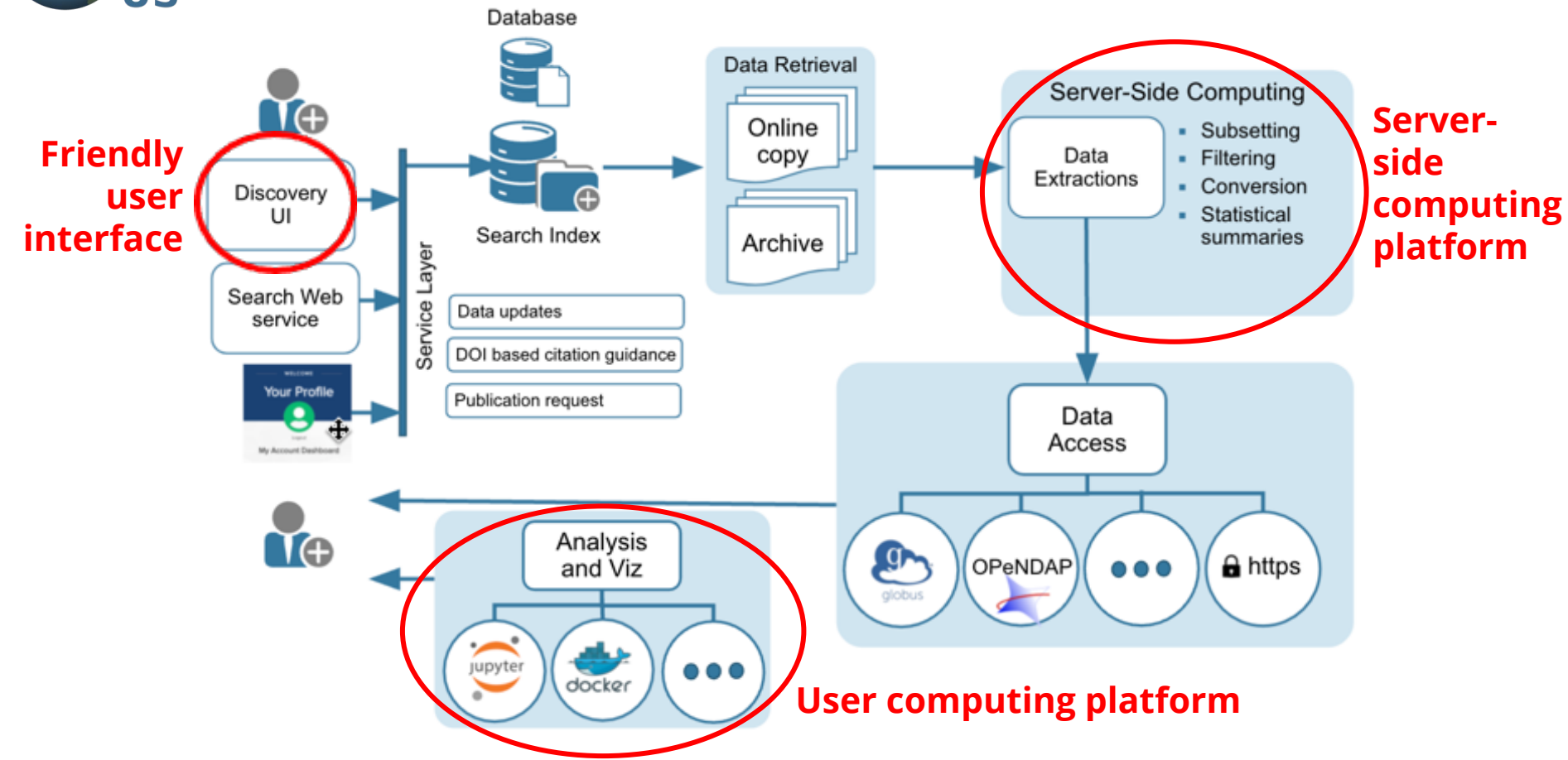
- Data Creators
  - Folks who submit CMIP decks and other modeling communities.
  - Users who analyse CMIP data and create derivative products (indices, tracks etc).
  - Folks who want to fuse observational products. ML etc...
- Data Consumers
  - Highly varied both in terms of **what** they want to do with the data and **how** they want to interact with it.
  - Many will have access to their own HPC and simply want Globus to suck the data out. Many will not and access to HPC is a barrier to doing science.
- Science Communities
  - There are traditional science communities very familiar with ESGF data and then there are untapped communities such as those in the resilience space. **We wish to increase data accessibility.**

Overarching goal: Remove barriers to doing science with ESGF-hosted data

ESGF US 2 User Computing and APIs

- User computing environment
  - *Jupyter-based computing platform*: Nimbus provides interactive development computational platform co-located with the data store.
  - *Containerized analysis environment*: The analysis environment will be containerized for execution on other HPC and cloud computing platforms.
  - *Streamlined access to ESGF data*: Will leverage and integrate ESGF central identity, search and transfers services/API to enable streamlined analysis workflows.
- Server-side APIs/services
  - Server-side subsetting: Extending ESGF REST API to allow requesting spatio-temporal subsets
  - Beyond files: API to provide data as Pandas/Xarray arrays that can be streamed for analysis
  - Summary statistics and visualizations: Standardized set of statistics and metrics computed on data orders
- Community tools and outreach
  - Rich set of sample Python notebooks/cookbooks for accessing and analyzing ESGF datasets.

ESGF 2 US Data Discovery Platform: Architecture



ESGF 2 US Summary of ESGF Services

Service	On-Premises	Cloud
Data	HTTPS, Globus, OPeNDAP	External ARCO
Indexing	Legacy Solr	Elastic-backed Globus Search
Identity, Auth	Specific providers participating in Globus	Globus Auth, EGI Checkin
Compute	JupyterHub in K8S clusters	*Option* to provision services
Web Application	Hosted	*Potentially* hosted

Thank you!

- ESGF Website: https://esgf.github.io/
- ESGF2-US on GitHub: https://github.com/esgf2-us/
- Contact: ames4@llnl.gov, forrest@climatemodeling.org