Open Science in the Cloud: Computing with Astrophysics Data Sets

HEASARC@GSFC

MAST@STScI

IRSA@IPAC/Caltech

Fornax@GSFC

Astrophysics flight mission data are already hosted in the commercial cloud

The commercial cloud is well-suited to store and make accessible the very large datasets from our current and future missions.

Many NASA Astrophysics datasets are now hosted in Amazon's AWS:

For almost all HEASARC data, see HEASARC Cloud Access

IRSA hosts Spitzer mosaic images, images and catalogs from WISE/NeoWISE/unWISE, and the OpenUniverse 2024 matched Rubin and Roman simulated skies: see IRSA Cloud Access

MAST hosts data from GALEX and Kepler/K2, and public data from current missions Hubble, TESS and JWST: see MAST Cloud Datasets and JWST Open Data. MAST will serve Roman mission data from the cloud.

Catalogs are stored in Apache Parquet format; they use HEALPix for efficient spatial searches, or HATS, a scheme developed in collaboration with Rubin/LINCC.

Download data to your local machine, or compute in the cloud??

NASA fornax

Some datasets are too big for your laptop, some software is too demanding

- **Data volumes** in the Astrophysics Archives **double roughly every 2 years:**
  - **Euclid** will release its first survey data in ~FY25 (Quick-Release1 ~30TB), with 2PB in FY26, and a further 6PB in FY28
  - **Roman** expects 1PB of level-2 science data (scan-by-scan) and level-3 data (co-added) after the first 6 months (~FY27), then another ~2PB per year
  - **Rubin Observatory (NSF-DoE)** will release 20TB/night from ~FY26, roughly 1.5PB per year, with a final data release of 15PB after 10 years of operation
- Multiwavelength astronomy requires **analyzing data sets jointly**; science often requires analyzing **data along with simulations** – pushes up the data volume further
- **Advanced data science tools** (e.g., AI/Machine Learning) can extract new knowledge, given adequate **computing resources close to the data**

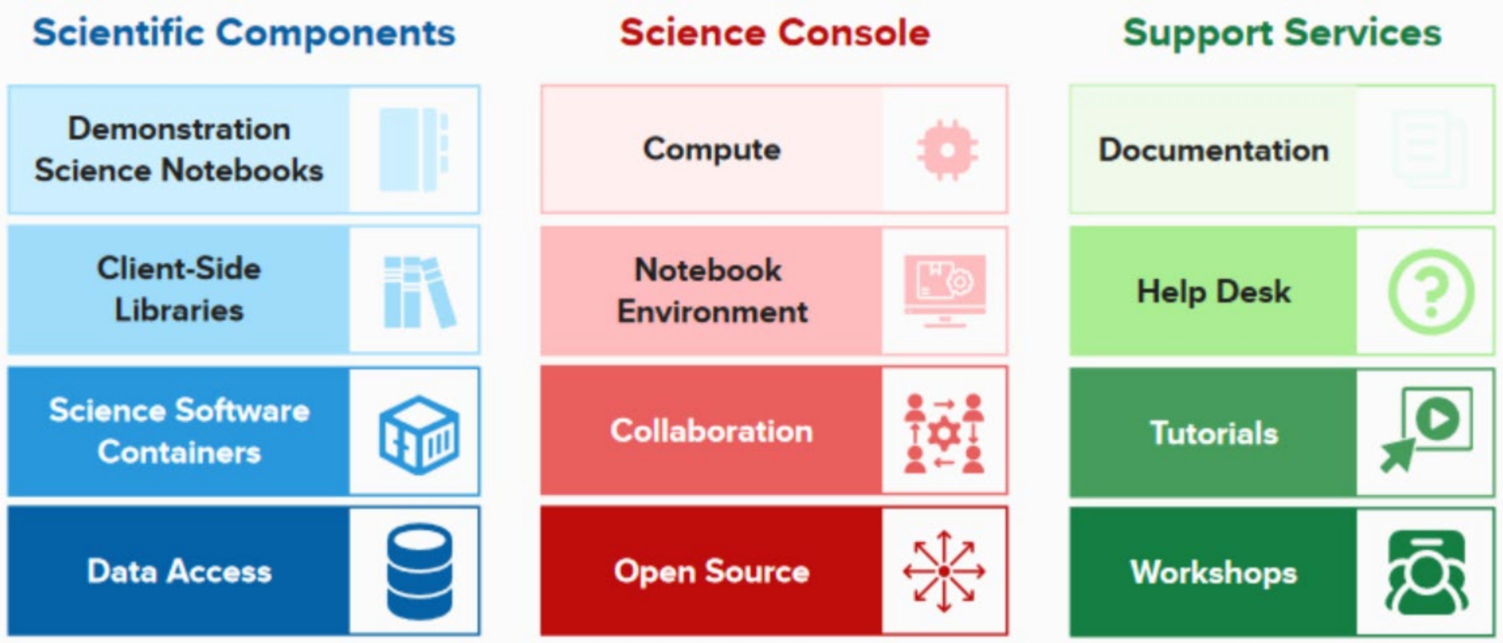
Fornax: compute in the cloud, proximate to cloud-held Astrophysics datasets

Fornax aims to provide science users with

- ➢ Access to a fixed allocation of cloud resources for storage, compute, etc, for free, from a standard web browser: no special hardware or software needed.
- ➢ Astronomy analysis tools, with notebooks that users can modify and run proximate to NASA and non-NASA data in the cloud; access to NASA Astrophysics data in the cloud and on-premises.
- ➢ Notebooks and documentation that allow users unfamiliar with particular techniques to learn those techniques and use them to analyze data.
- ➢ Ability for users to upload (limited amounts of) their own data or code to run, and export results
- ➢ Ability to form collaboration groups that share notebooks, data files, etc, while keeping them private to your collaboration.

NASA Astrophysics is developing Fornax in the Amazon (AWS) cloud

What will Fornax provide for users of Astrophysics mission data?



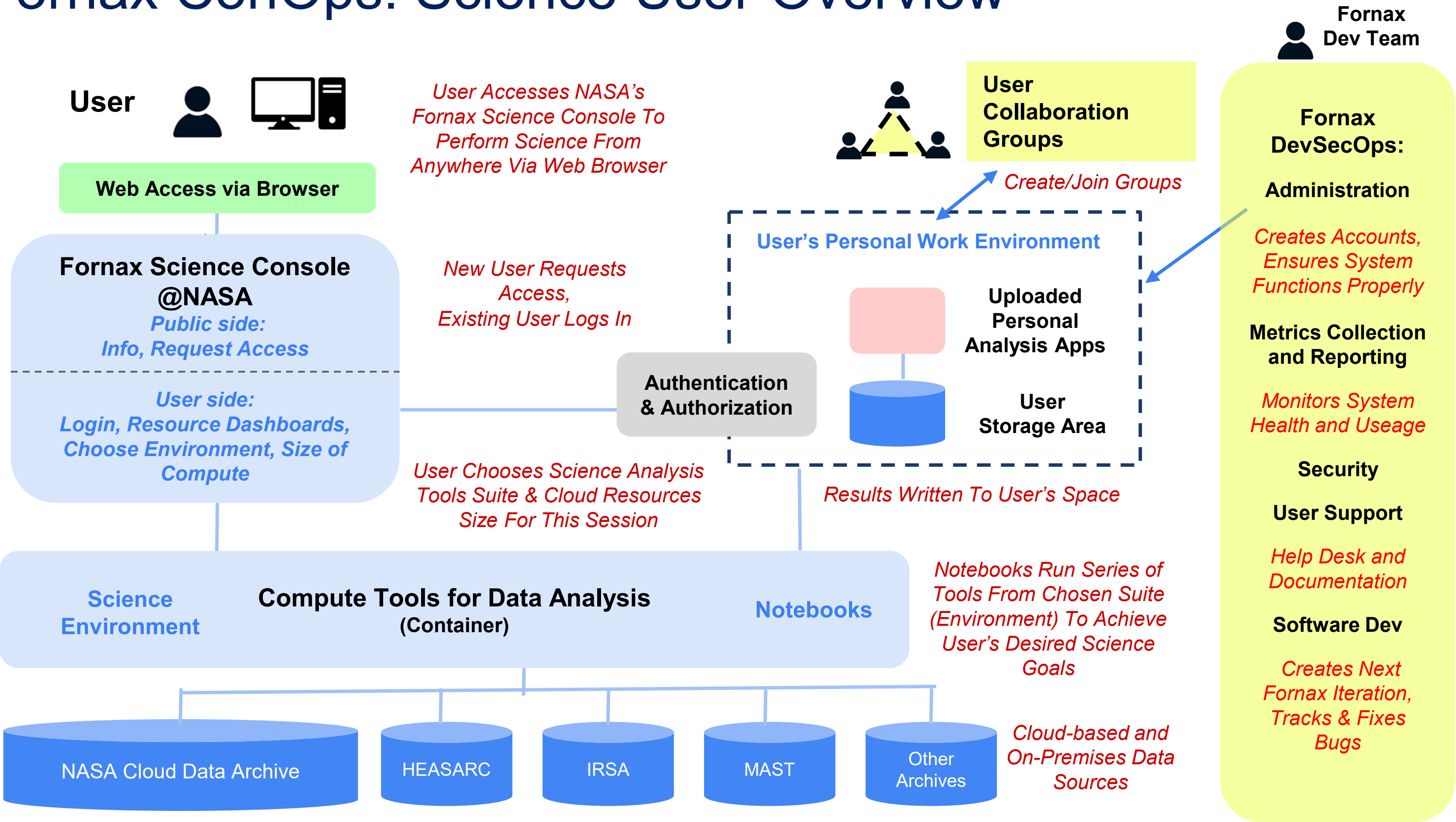
Scientific Components	Science Console	Support Services
Demonstration Science Notebooks	Compute	Documentation
Client-Side Libraries	Notebook Environment	Help Desk
Science Software Containers	Collaboration	Tutorials
Data Access	Open Source	Workshops

The archives HEASARC, IRSA and MAST curate the datasets, whether cloud-held or on-premises

Science users interact directly with these Fornax elements:

- **Science Notebooks**: library of runnable tutorial notebooks written in Python to perform science data analysis tasks. Many already exist, some will be developed or modified for Fornax. Users can modify the notebooks for their science analysis.
- **Containerized Science Environments**: containers preinstalled with major astrophysics data analysis software, with options for different contexts (e.g., high energy, optical photometry).
- **Client-side libraries**: open source tools (e.g., astroquery or image cutout tools) that can make use of cloud data and services.
- The **Fornax Science Console – ForSC** ("the Force"): open source infrastructure-as-code to deploy a Jupyter-based user interface for these cloud resources. NASA's deployment will provide users with a fixed allowance of cloud-computing resources without charge, proximate to Astrophysics mission data in the cloud.

Fornax ConOps: Science User Overview



User

Web Access via Browser

User Accesses NASA's Fornax Science Console To Perform Science From Anywhere Via Web Browser

Fornax Science Console @NASA

Public side: Info, Request Access

User side: Login, Resource Dashboards, Choose Environment, Size of Compute

New User Requests Access, Existing User Logs In

Authentication & Authorization

User Chooses Science Analysis Tools Suite & Cloud Resources Size For This Session

User Collaboration Groups

Create/Join Groups

User's Personal Work Environment

Uploaded Personal Analysis Apps

User Storage Area

Results Written To User's Space

Science Environment

Compute Tools for Data Analysis

(Container)

Notebooks

Notebooks Run Series of Tools From Chosen Suite (Environment) To Achieve User's Desired Science Goals

NASA Cloud Data Archive

HEASARC

IRSA

MAST

Other Archives

Cloud-based and On-Premises Data Sources

Fornax Dev Team

Fornax DevSecOps:

Administration

Creates Accounts, Ensures System Functions Properly

Metrics Collection and Reporting

Monitors System Health and Useage

Security

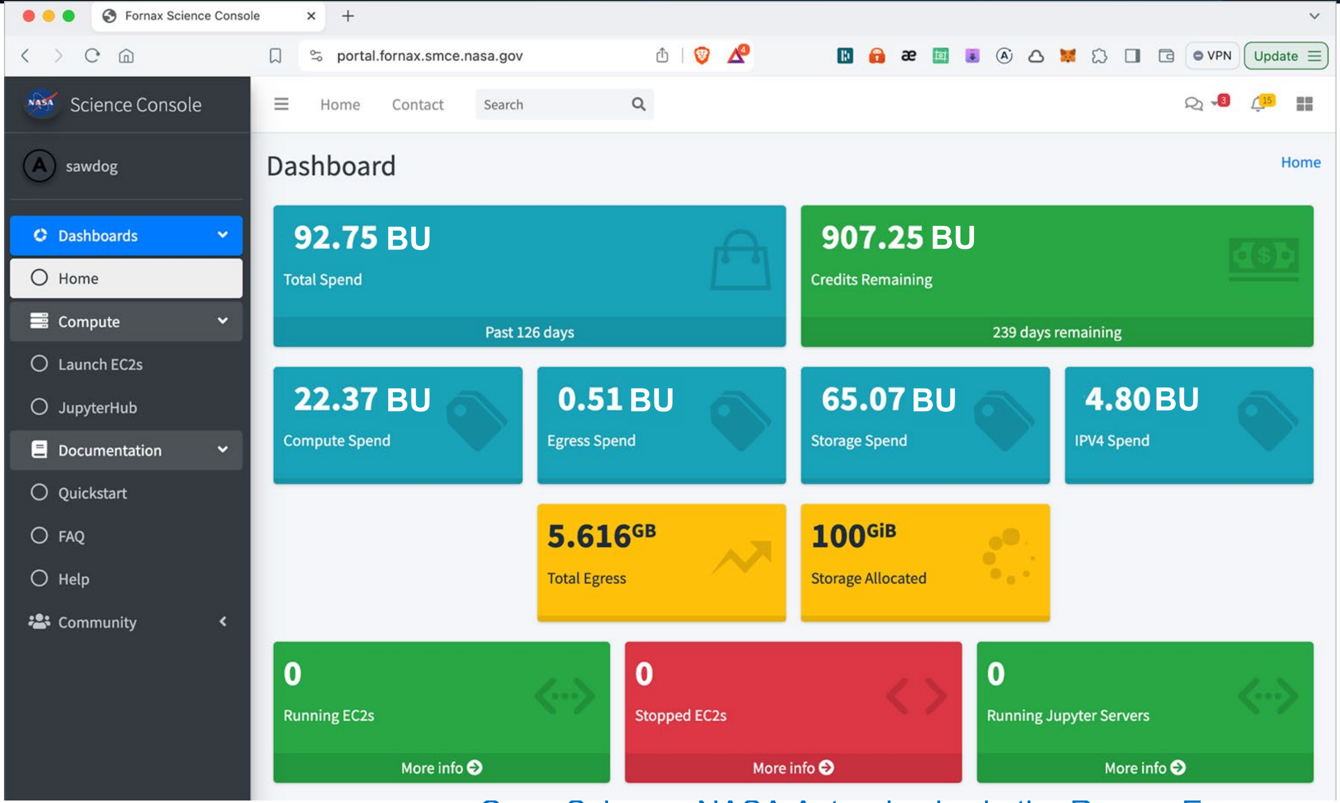
User Support

Help Desk and Documentation

Software Dev

Creates Next Fornax Iteration, Tracks & Fixes Bugs

Users will see their resource usage on a dashboard like this:



Dashboard

Home

92.75 BU
Total Spend
Past 126 days

907.25 BU
Credits Remaining
239 days remaining

22.37 BU
Compute Spend

0.51 BU
Egress Spend

65.07 BU
Storage Spend

4.80 BU
IPV4 Spend

5.616GB
Total Egress

100 GiB
Storage Allocated

0
Running EC2s
More info

0
Stopped EC2s
More info

0
Running Jupyter Servers
More info

Computing in the cloud is not like working on your laptop

To err is human, to make a real mess requires a computer...

If you start your cloud session and then go to lunch before beginning a compute task, you are wasting resources – could be a lot!

Downloading (egress) generally costs resources. The cloud stores huge volumes of data, so you easily burn up your resources by working inefficiently or downloading more than you need. Use cloud-optimized libraries for streaming data into your session, manage wisely what data you do move.

Don't assume that a Jupyter notebook developed on your laptop will perform well in the cloud. Workflows requiring significant data I/O, memory management, or CPU parallelization are very different in the cloud. Tutorial notebooks (next slide) point the way!

Try with a small sample first before tackling your full problem. Don't be surprised if something goes wrong. Any new combination of sample selection, which archives are called and what parameters are used, runtime environment, machine CPU and RAM capabilities, network bandwidth, etc. will present new reasons for the code to fail. Observe how the code performs, diagnose a problem, and adapt the input parameters, machine size, etc.

And remember to **shut down your running cloud instances when you are done working** – don't just close the browser window and walk away!

Example and tutorial notebooks

- Working with TESS data in the cloud: TESS intro using TIKE
- IRSA cloud access tutorial introduction (Parquet basics using AllWISE – you can run this on a laptop)
- Light Curve Generator, to run on Fornax or elsewhere:
  - Can query S3 unWISE parquet. 500k light curves in 4 hours (<< 100 day est. for original code to query FITS files at NERSC)
- AllWISE Source Catalog Examples on IRSA's github
- 5-minute Quick Start with LSDB/HATS on lincc-frameworks github – this was run on Fornax as a demo for the IVOA meeting in November 2024
- Cross-match catalogs from ZTF and Pan-STARRS using LSDB (a useful package for large-catalog cross-match): ZTF-Pan-STARRS X-match
- Matching up galaxies in the SDSS MaNGA survey with HST images to make combined images: MaNGA-HST X-match & images
- Run the JWST NIRCam imaging pipeline: NIRCam imaging

Can I get an account right now?

- Fornax is still developing user management tools and security
- Thus far Fornax hosts "internal" beta users only; we plan to onboard other beta users starting in FY 2026.
- Info coming soon at pcos.gsfc.nasa.gov/Fornax/

Backups/unused slides