The Fornax Initiative

A NASA Astrophysics Science Platform

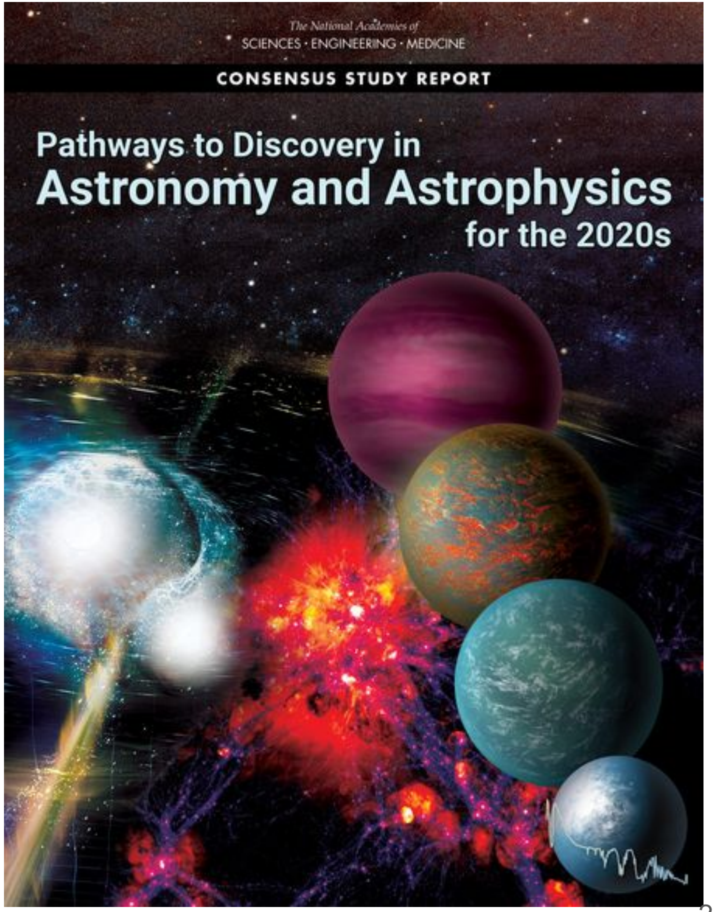
Tess Jaffe presenting on behalf of the team:
Patrick Coronado[1] (Project Manager), Tess Jaffe[2] (Project Scientist), Francesca Civano[3] (Deputy Project Scientist), Vandana Desai[4], Steven Groom[4], Jonathan Hargis[5], George Helou[4], Josh Peek[5], Andy Ptak[2], Gail Rohrbach[1], et al.
[1]Astrophysics Projects Division, NASA/GSFC
[2]High Energy Astrophysics Science Archive Research Center, NASA/GSFC [HEASARC]
[3]Astrophysics Projects Division, Physics of the Cosmos Program Office, NASA/GSFC
[4]NASA/IPAC Infrared Science Archive, Caltech [IRSA]
[5]Mikulski Archive for Space Telescopes, STScI [MAST]

The challenge

"The importance of joint analysis of observations from different facilities and wavelengths, and of sophisticated archiving with associated science platform tools, will grow dramatically over the next decade."

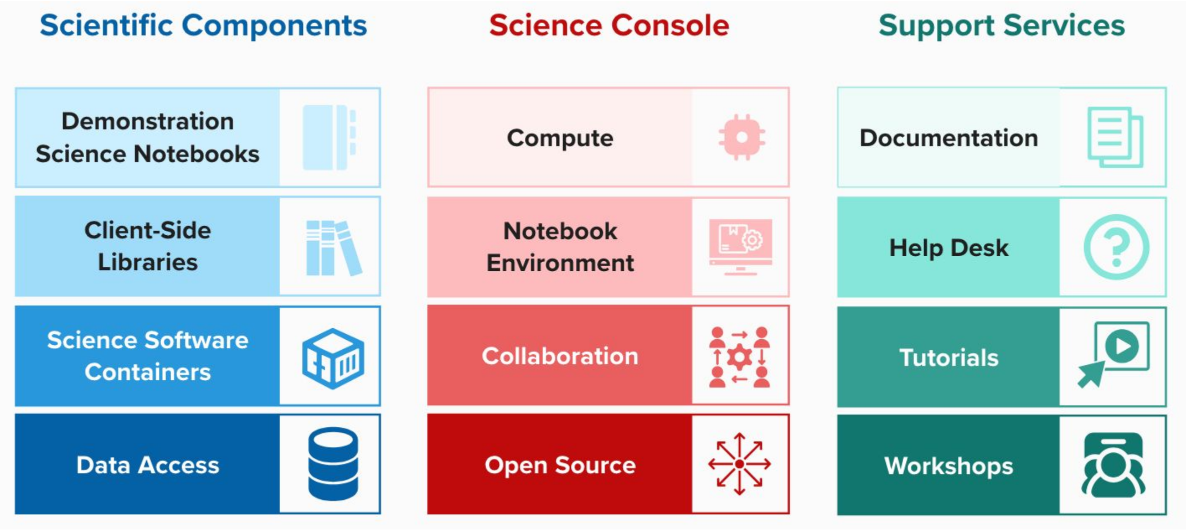
The Astrophysics Division sees a common need in our overlapping communities for:

- multi-archive analyses on large data sets ("big data");
- complex analyses that require a lot of compute ("big compute");
- collaboration tools and runnable examples for non-experts ("big community").



The Fornax Initiative

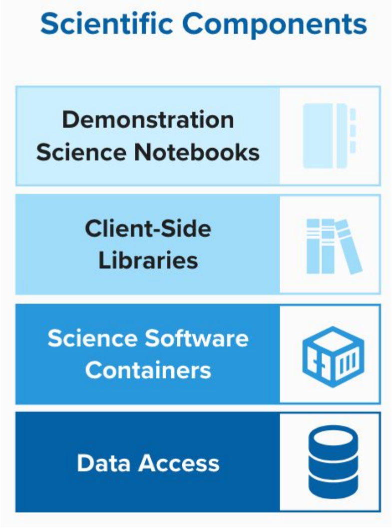
Fornax is a NASA Astrophysics Archives effort to collaboratively create cloud systems, cloud software, and cloud standards for the astronomical community.



Fornax users: early career scientists without analysis experience; any scientist without adequate computational facilities; any scientist wishing to analyse large amounts of data proximate to the compute in the cloud; collaborators at different institutions wishing to share a computational environment; etc.

The Fornax Initiative: Scientific Components

The astrophysics-specific elements required to enable science in the cloud, including Python notebooks that demonstrate access to cloud-hosted NASA mission data, curated astrophysics software environments, and cloud-native services to support common astronomy workflows.



All these components can be used *à la carte*:

- Notebooks are portable.
- Software is open source.
- Software environments will be in public container registries.
- Data can be accessed from anywhere.

The Fornax Initiative: Scientific Components

Python client for data search and retrieval:

```
pyvo.utils.download_file(record,'aws')
```

Under the hood:

- Service layer returns *multiple* access options.
- Client layer selects the right access method for cloud versus on-prem.

Cloud-optimized large catalog cross matching (collaboration with [LINCC Frameworks]) to enable workflows like:

```
img = gaia
      .query("pm > 10")
      .crossmatch(ztf)
      .joint(ztf_sources)
      .for_each(varstar_classify)
      .query("pRRLy > 0.95")
      .skymap()
hp.mollview(img)
```

See https://lsdb.readthedocs.io/.

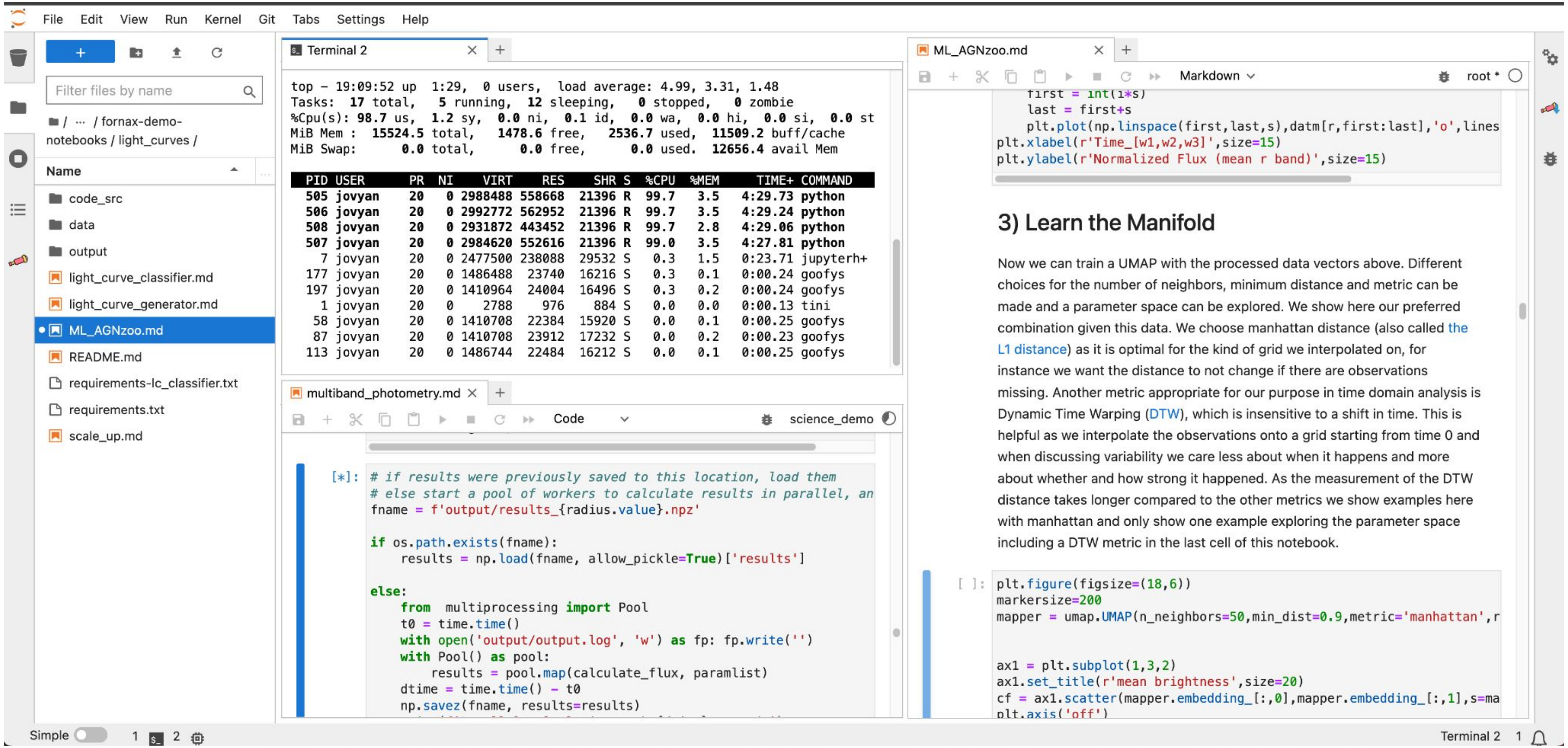
Or partial FITS reads out of S3 in astropy:

```
from astropy.io import fits

# URI of a 213 MB FITS file hosted in a free Amazon S3 cloud storage bucket
uri = "s3://stpubdata/hst/public/j8pu/j8pu0y010/j8pu0y010_drc.fits"

# Download the primary header
with fits.open(uri) as hdul:

    # Download a single header
    header = hdul[1].header

    # Download a single image
    mydata = hdul[1].data

    # Download a small cutout
    myslice = hdul[2].section[10:12, 20:22]
```

The entire 213 MB file is never downloaded. Only the chunks you want are transferred.

The Fornax Initiative: Scientific Components

Notebooks demonstrating real science use cases:

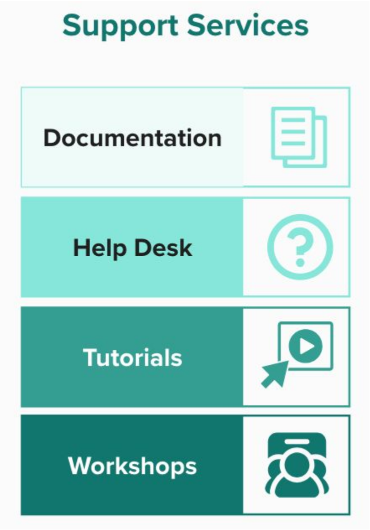


The Fornax Initiative: Support Services

A program of engagement with the astronomical community, including a Helpdesk and training opportunities

User documentation under development at:
https://fornax-navo.github.io/fornax-documentation/
(We are working in an open source and transparent way

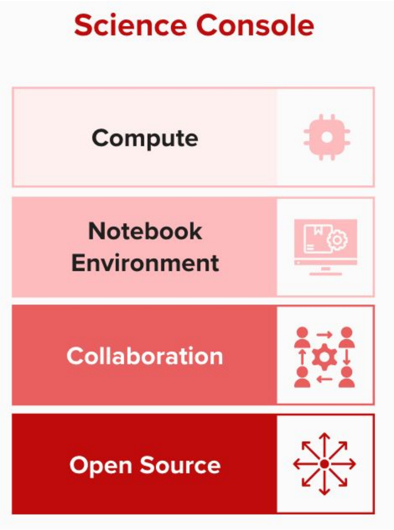
Workshops coming in the next few years to a meeting near you.



The Fornax Initiative: Science Console

A web-based application that users log into for access to cloud computing, data storage, and interactive data analysis in JupyterLab.

- Common infrastructure elements.
- Collaborating with other platform projects.
- Using whatever already exists where possible.



- Open sourcing whatever we develop.
- Deployable elsewhere,
    - i.e. a university department could deploy their own Fornax Science Console on their own AWS account.

ConOps: Science User Overview-1

User

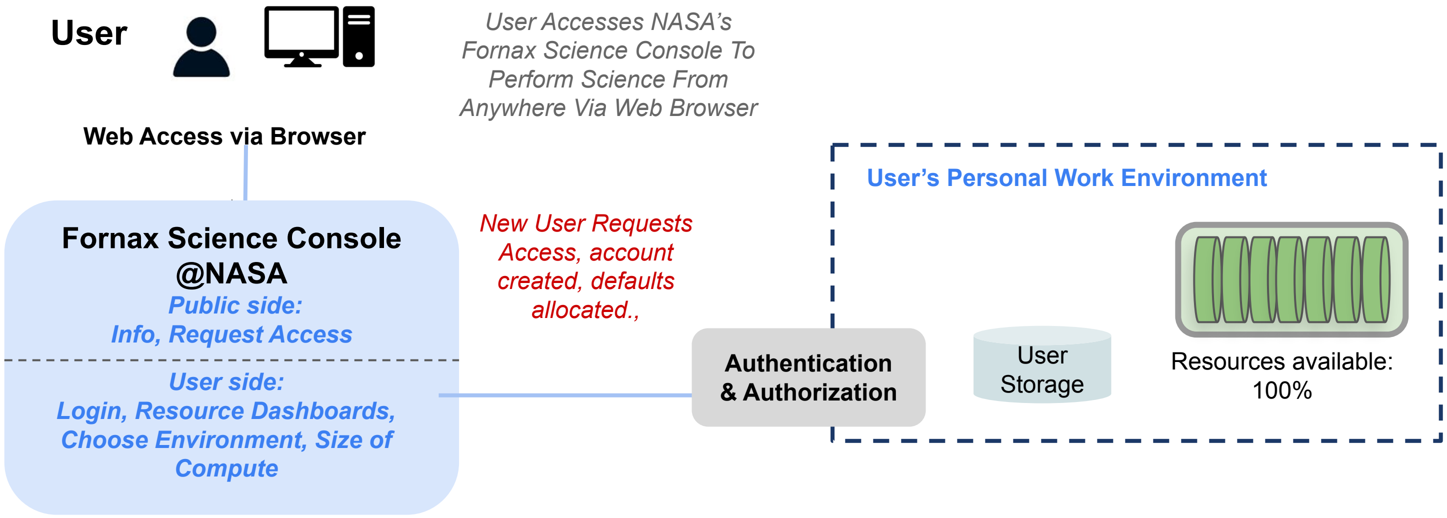
User Accesses NASA's Fornax Science Console To Perform Science From Anywhere Via Web Browser

Web Access via Browser

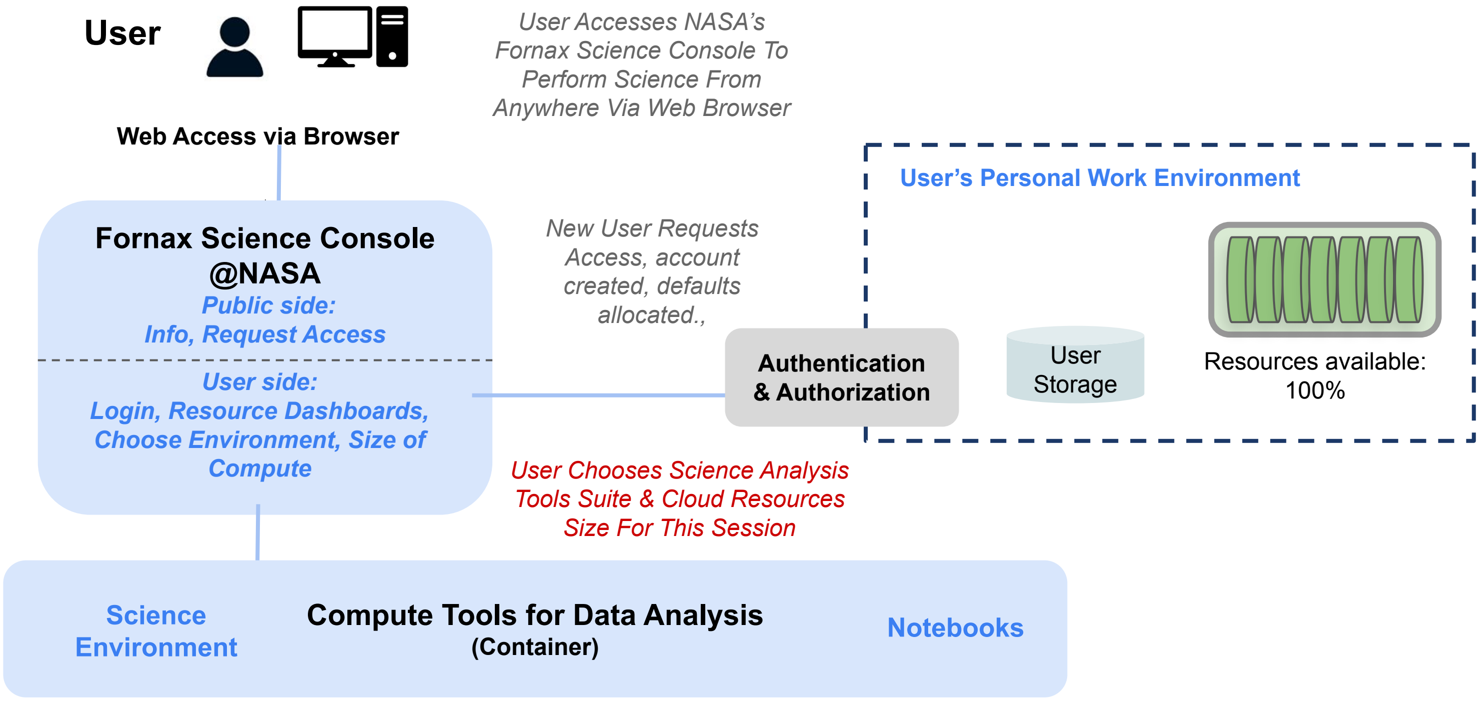
Fornax Science Console @NASA

Public side:
Info, Request Access

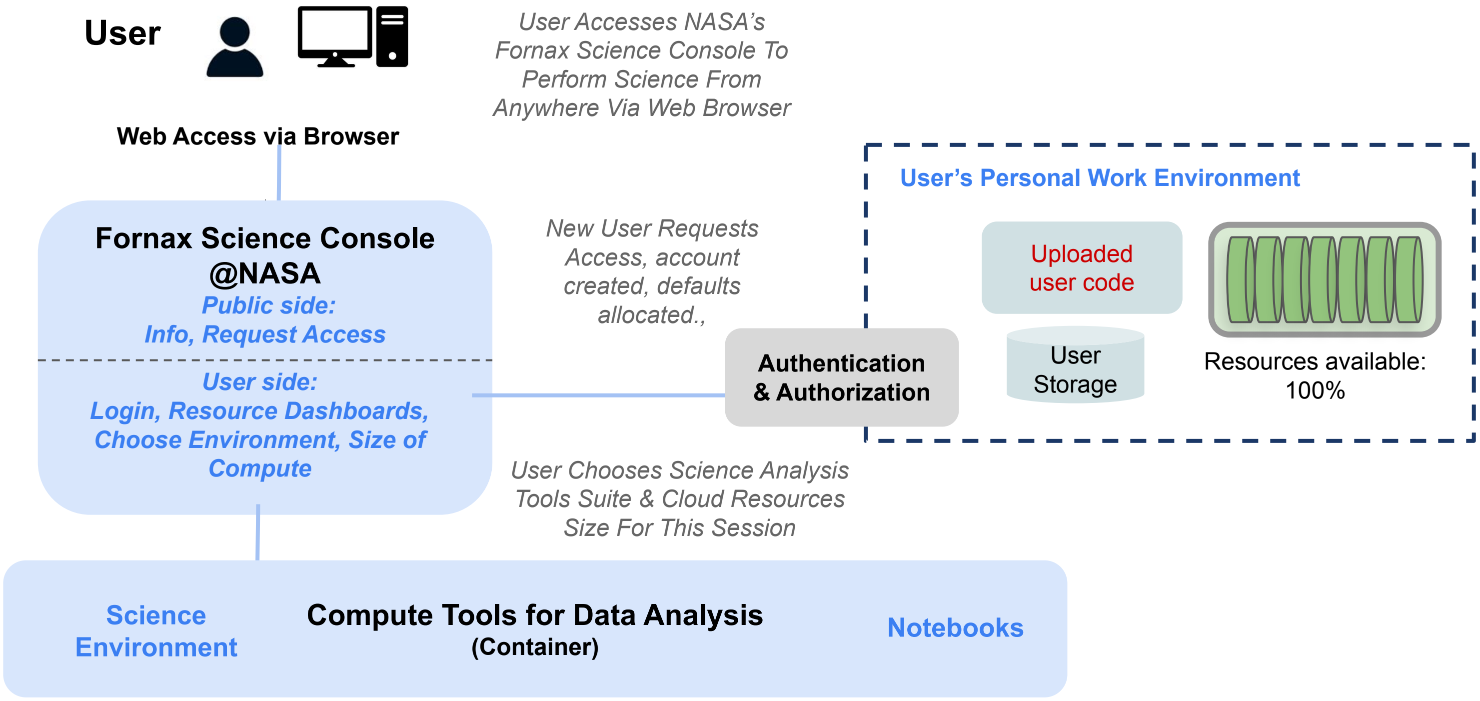
ConOps: Science User Overview-1



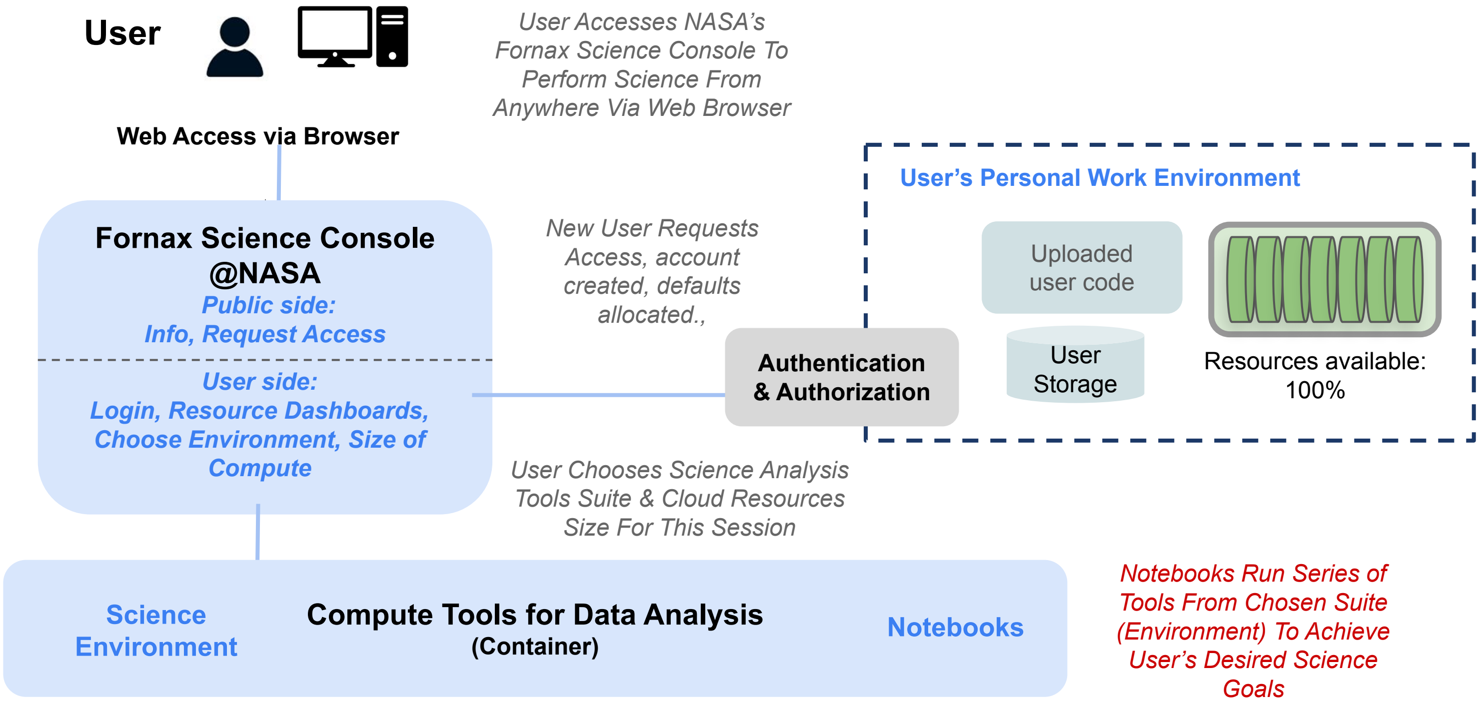
ConOps: Science User Overview-1



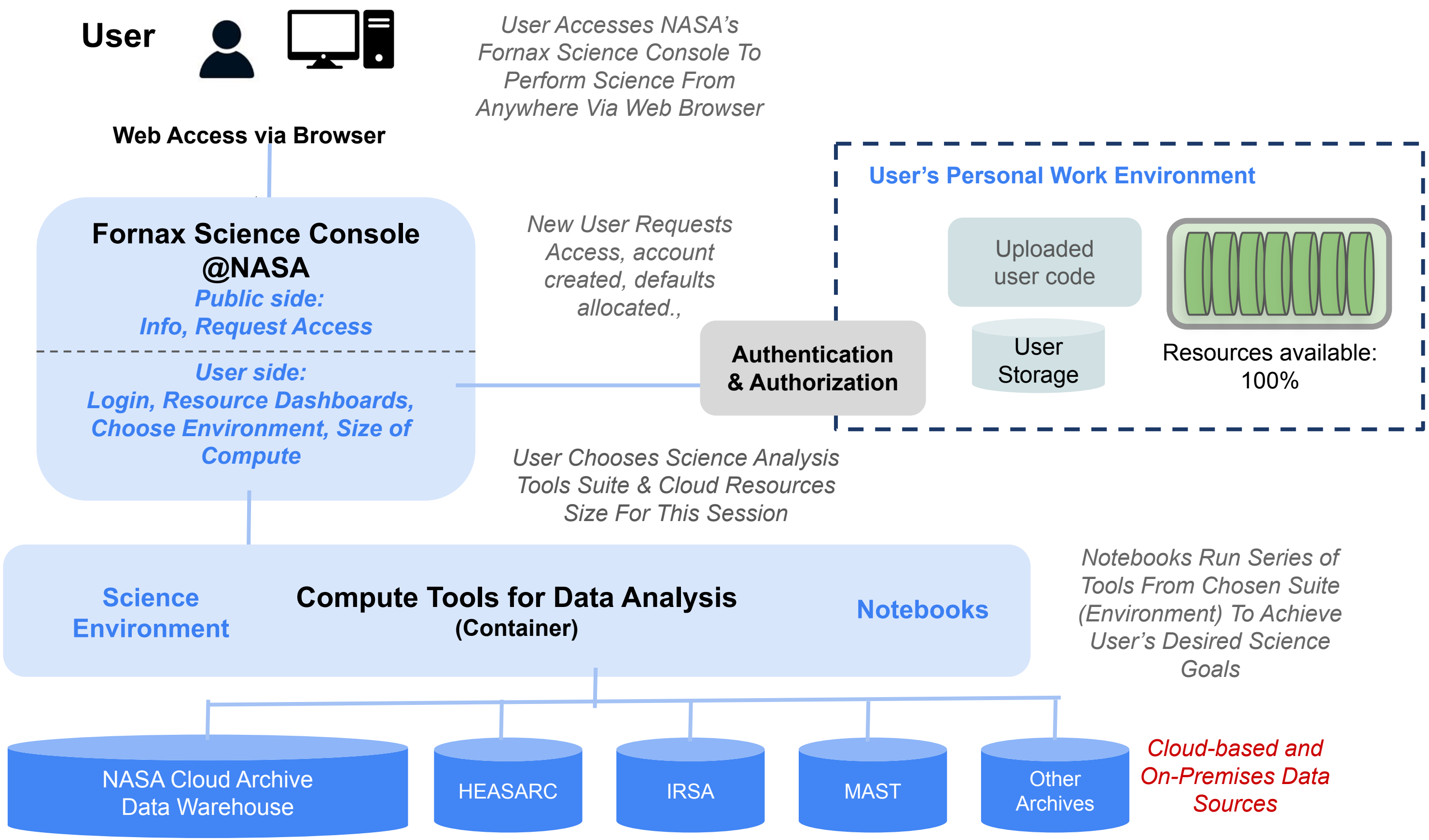
ConOps: Science User Overview-1



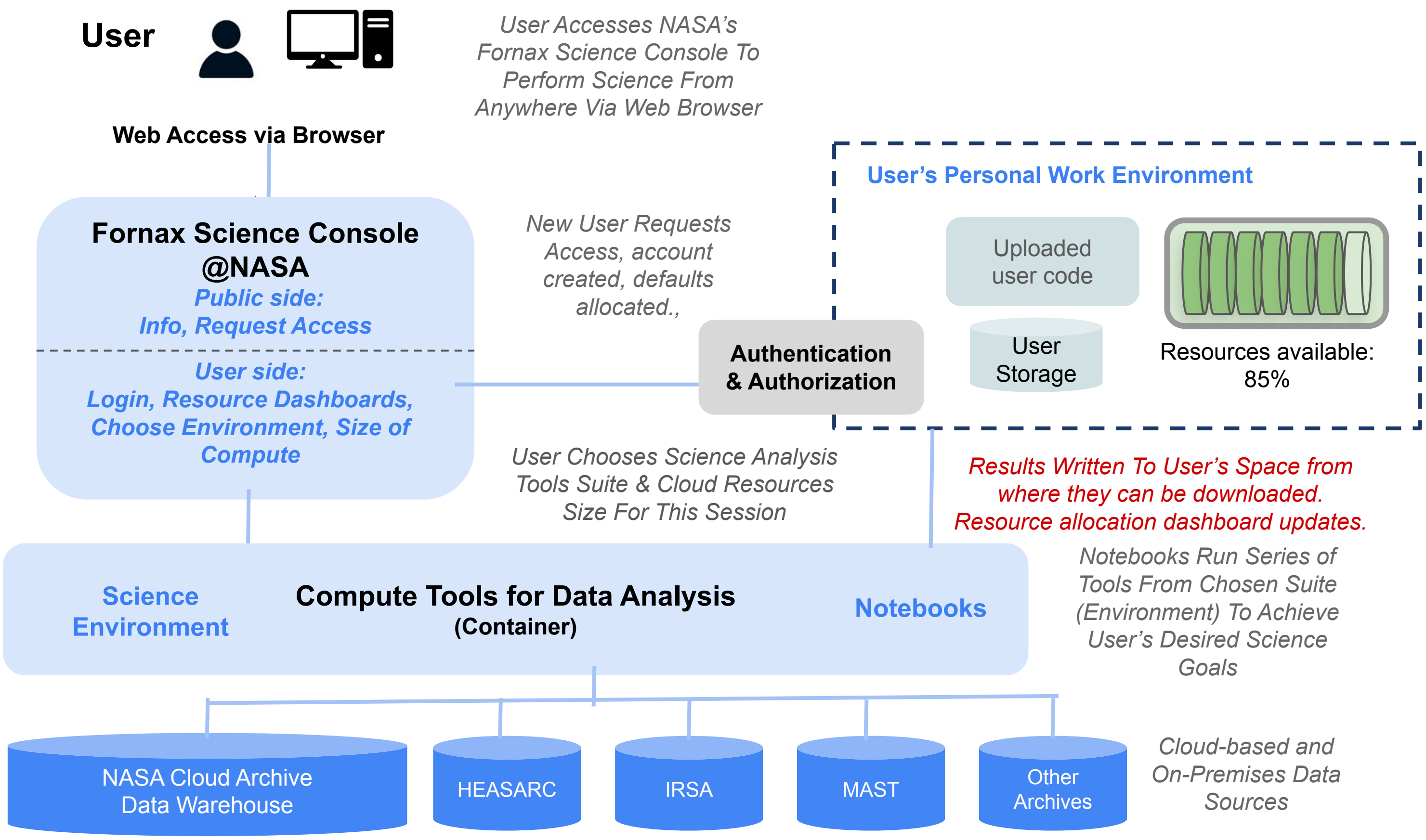
ConOps: Science User Overview-1



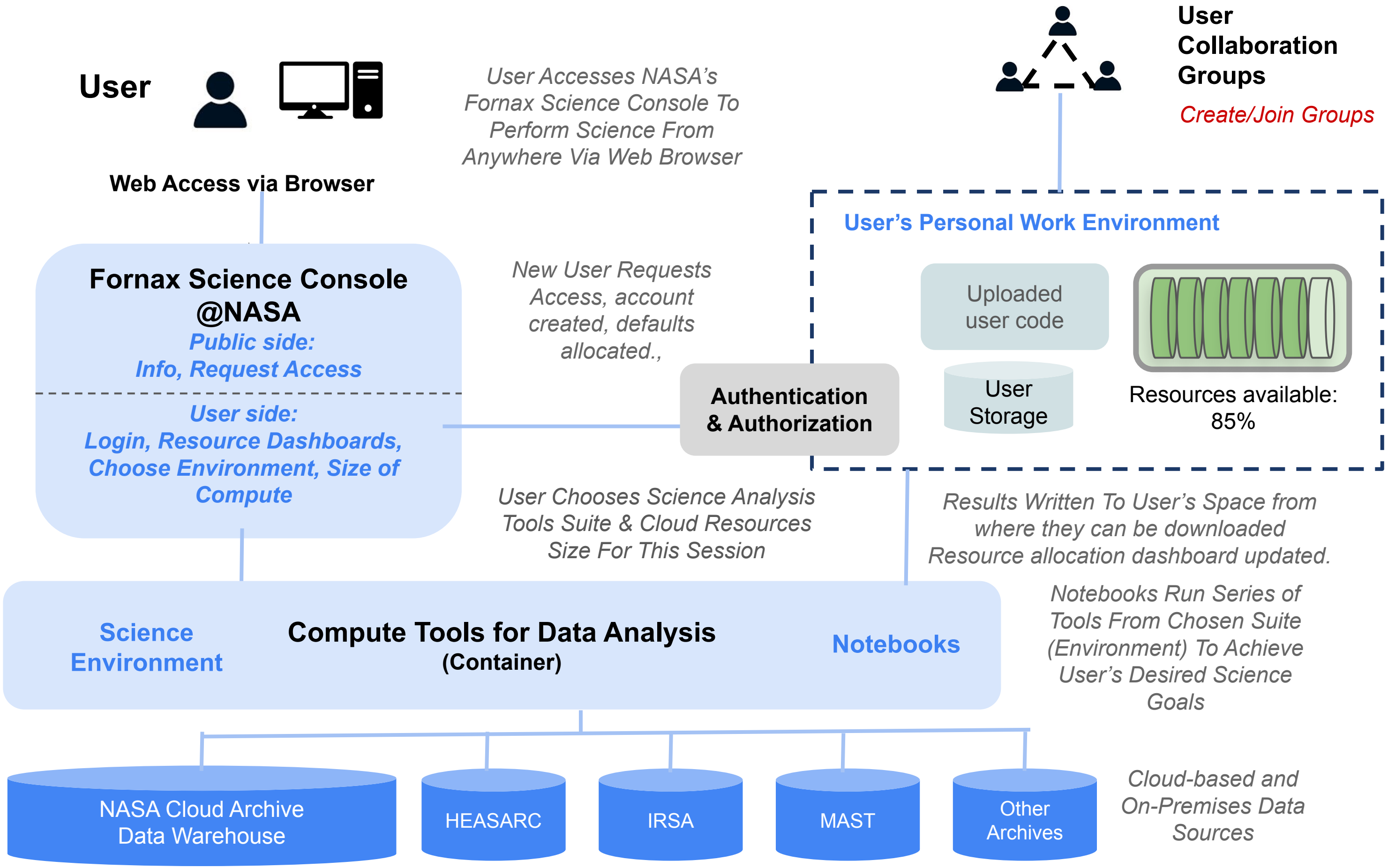
ConOps: Science User Overview-1



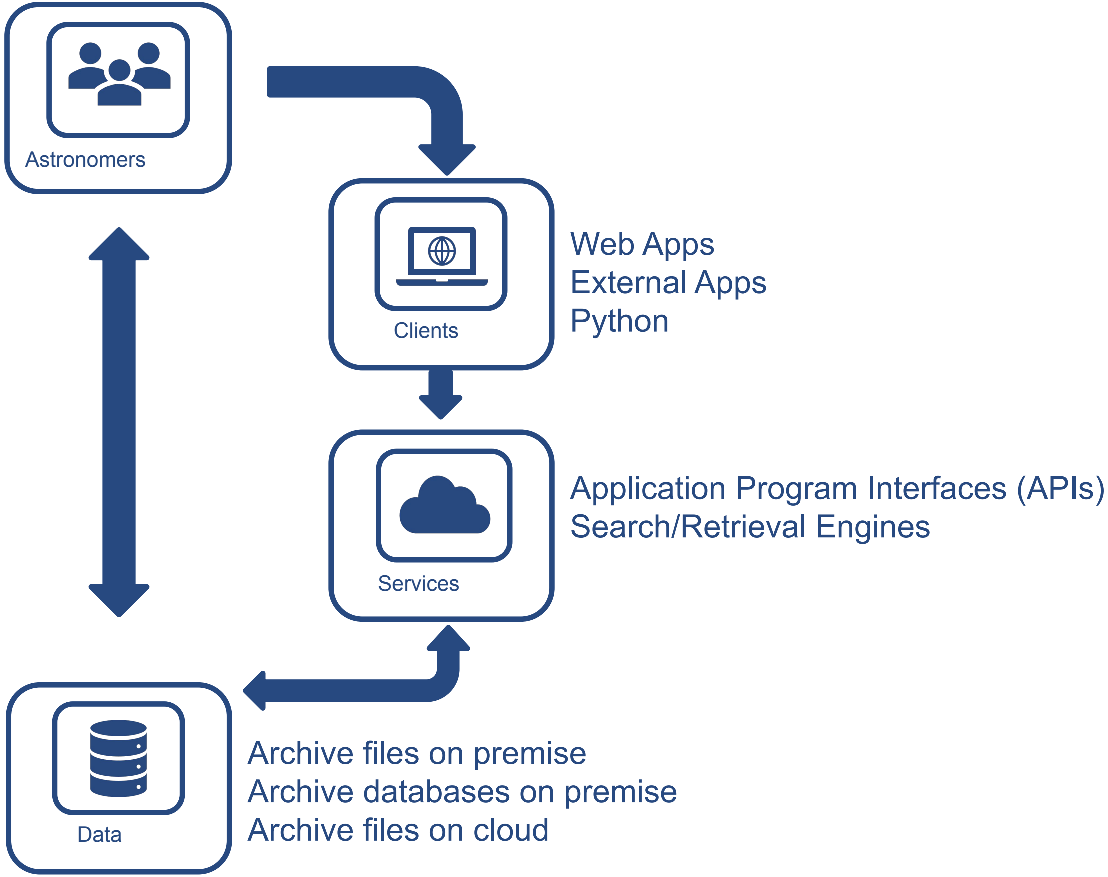
ConOps: Science User Overview-1



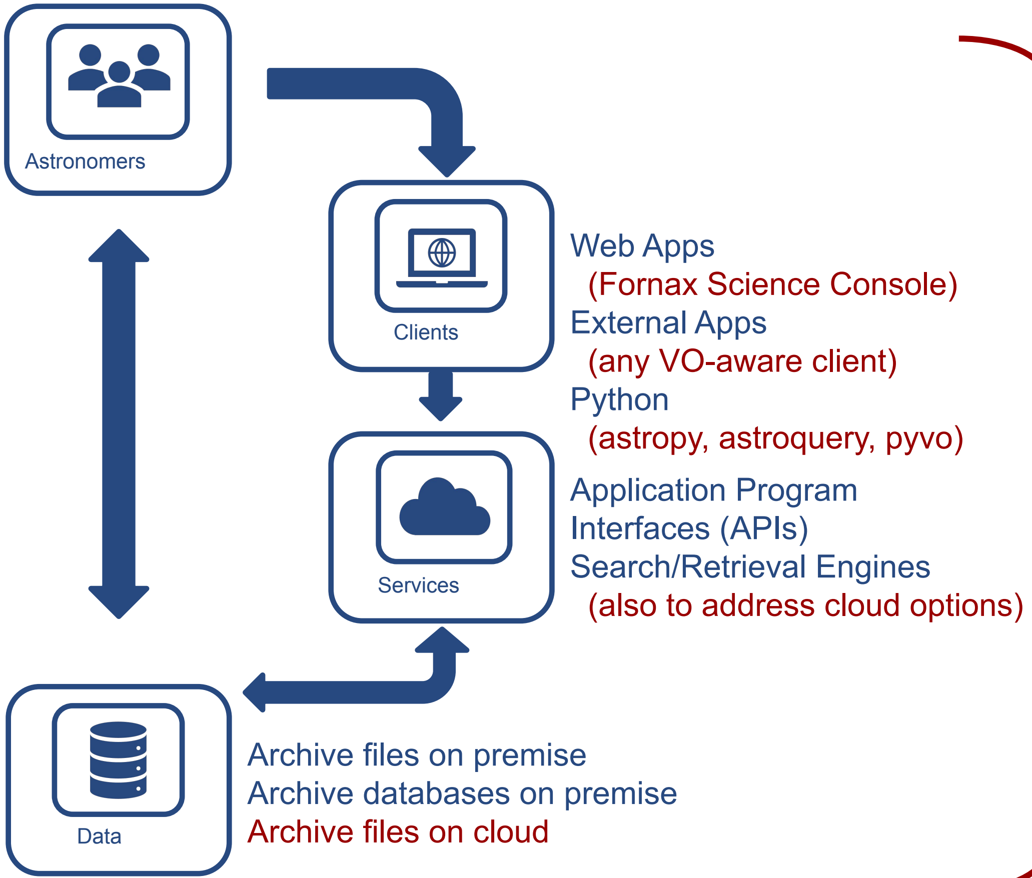
ConOps: Science User Overview-1



Generalized archival analysis architecture:



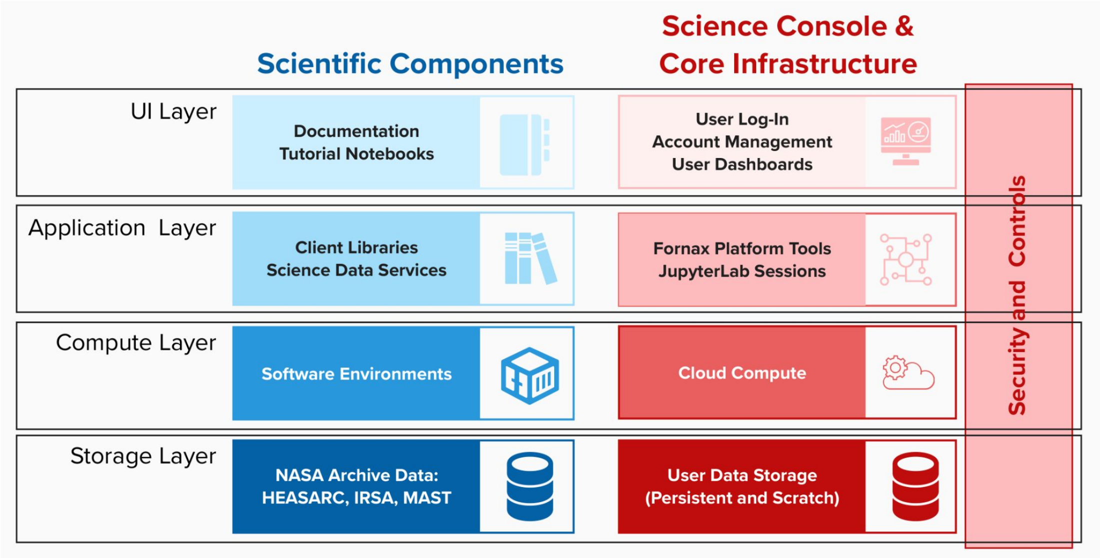
Bottom line:



This workflow is getting more options using cloud computing. Stay tuned for how to facilitate your science with Fornax.

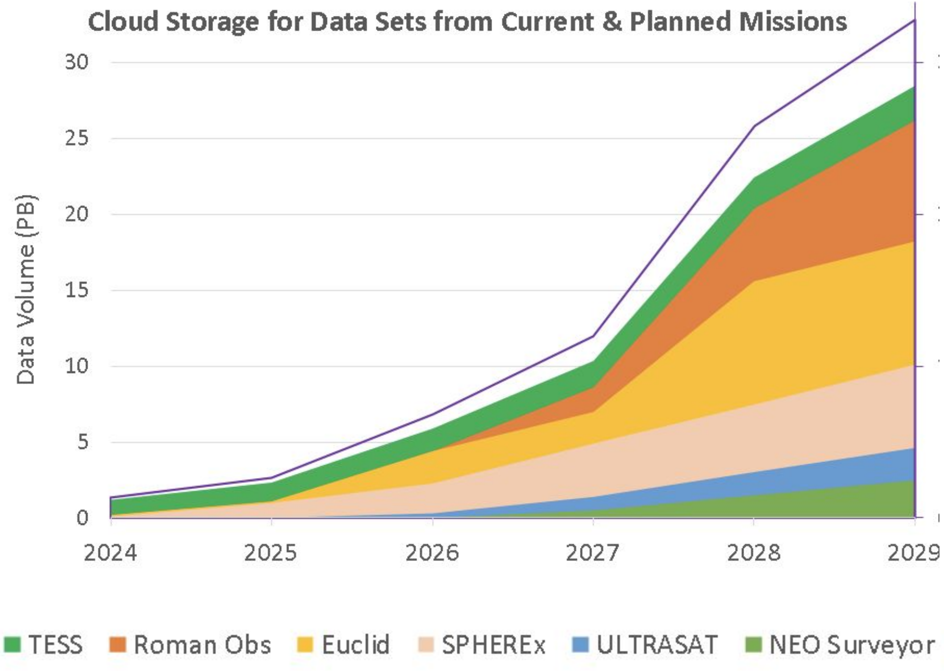
https://pcos.gsfc.nasa.gov/Fornax/

Fornax architecture



Evolution of Data (Storage) Layer: Cloud Storage

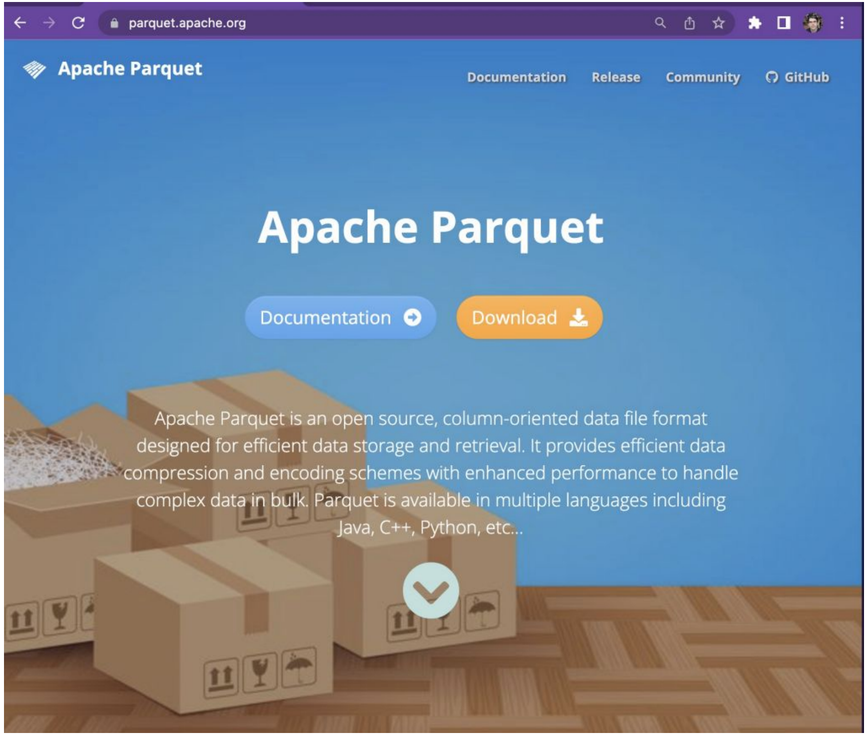
- All NASA Astrophysics Mission Archives now have popular data sets *copied* to the cloud
  - HEASARC on AWS
  - IRSA on AWS
  - MAST on AWS
- In the future, we will increase the volume of data that is *only* served from the cloud



Data

Evolution of Data (Storage) Layer: File Formats

- **Images** stored as FITS files, as on prem
- **Catalogs**
    - On-prem databases currently serve catalog data
    - In cloud, large catalogs will be served in analysis-ready Parquet format (working name: HiPSCat)
    - HiPSCat partitioning scheme to support cross-matching currently being tested across archives/missions



[See talk by Mario Juric at IVOA interop meeting May 2023]

Data

Current State of Services Layer

The NASA Astrophysics Archives are part of the NASA Astronomical Virtual Observatories (NAVO) and have adopted a uniform set of Application Program Interfaces (APIs) standardized by the International Virtual Observatory Alliance (IVOA):

- Images: Simple Image Access (SIA)
- Spectra: Simple Spectral Access (SSA)
- Catalogs:
  - Simple Cone Search (SCS)
  - Table Access Protocol (TAP)
- Data Products: ObsTAP and DataLink

Evolution of Services Layer: Support for cloud-hosted files

- All APIs and underlying data services must be updated to provided on-prem and/or cloud pointers to data.
- New IVOA standards must be created to do this!

Services

Evolution of Client Layer: Support for cloud-hosted files

PyVO

- an affiliated astropy package.
- lets you find and retrieve astronomical data available from archives that support standard IVOA service protocols.
- needs to be updated to include pointers to data in the cloud.

Client libraries:

```
pyvo.utils.download_file(record,'aws')
```

Under the hood:

- Service layer returns multiple access options.
- Client layer selects the right access method for cloud versus on-prem. (TBD)

Clients

Evolution of Client Layer: Support for subsets of cloud-hosted FITS

- It’s common to only need a portion of a large FITS file.
- Downloading an entire FITS file can take a long time, and can also run into memory limitations.
- Modifications were made to popular astronomy Python package astropy to provide an API that will work the same whatever the storage backend.

```python
from astropy.io import fits

# URI of a 213 MB FITS file hosted in a free Amazon S3 cloud storage bucket
uri = "s3://stpubdata/hst/public/j8pu/j8pu0y010/j8pu0y010_drc.fits"

# Download the primary header
with fits.open(uri) as hdul:

    # Download a single header
    header = hdul[1].header

    # Download a single image
    mydata = hdul[1].data

    # Download a small cutout
    myslice = hdul[2].section[10:12, 20:22]
```

The entire 213 MB file is never downloaded. Only the chunks you want are transferred.

Clients

Evolution of Client Layer: support for large catalogs in Parquet format

[See talk by Mario Juric at IVOA interop meeting May 2023]

LSDB: Python Analytics for HiPSCat

- LSDB: Large Survey Database
- Enable Pandas-like analysis on trillions of observations with thousands of cores
- Build on existing tools: Dask (looking at Ray).
- Full HiPSCat awareness: spatial queries, cross-matching, timeseries, multi-dataset joining.
- Very much in pre-alpha/prototype phase; expect usable alphas in the next few weeks

```
img = gaia
     .query(“pm > 10”)
     .crossmatch(ztf)
     .join(ztf_sources)
     .for_each(varstar_classify)
     .query(“pRRLy > 0.95”)
     .skymap()

hp.mollview(img)
```

LSDB target APIs: The API center science. Multi-processing, autoscaling, fail-over, etc. are all implicit. Good user experience.

Wyatt et al. (2023)
https://github.com/astronomy-commons/lsdb

Clients