

# Reproducible remote pipelines

The EVN archive portal

Aard Keimpema

[keimpema@jive.eu](mailto:keimpema@jive.eu)

# JIVE and the EVN

- JIVE is located in the Netherlands
- Technical operations and user support for the European VLBI Network (EVN)
- The EVN is an *Open Skies* facility
- All EVN data is public after 1 year
- Archive contain ~180 TB raw astronomical data



The European VLBI Network



Calibration  
&  
Imaging



- De facto standard data reduction package for radio astronomy
- Under active development since 1990s
- Mostly C++, but has python bindings to all tasks
- VLBI support developed at JIVE in ESCAPE WP3

The screenshot displays the CASA software interface. On the left, a 'Logging window' shows a list of messages including iteration stepping and plotting of unflagged points. Below it, an 'IPython shell' displays the CASA startup sequence, including the IPython version and the execution of the `plotms` command. On the right, a 'Plotting tool (Qt based)' window titled 'PlotMS' shows four subplots of 'Gain Amp vs. Channel' for antennas ea01@W09, ea02@E02, ea03@E09, and ea04@W01. The plots show gain amplitude versus channel number, with a control panel for file selection and averaging options.

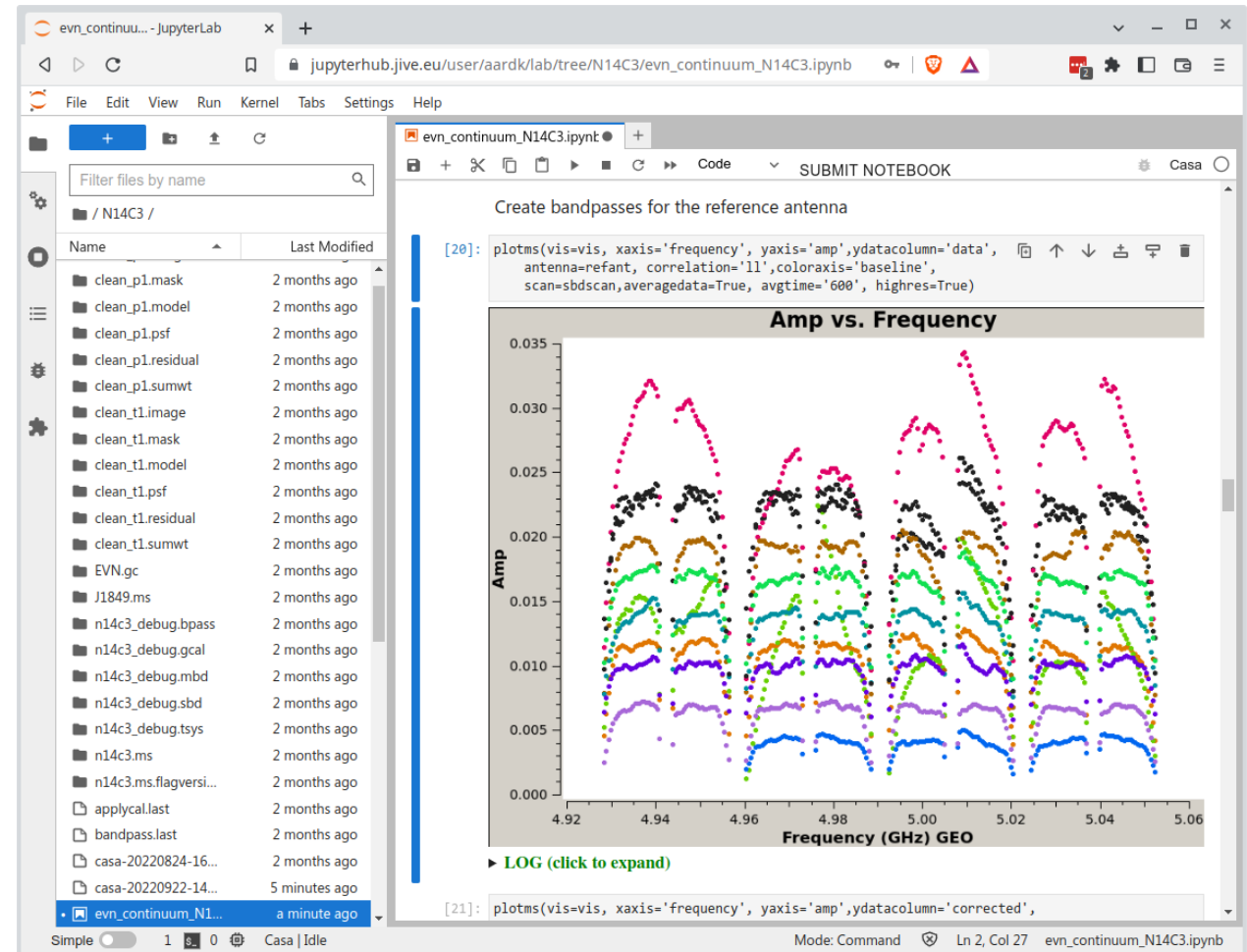
Screenshot of CASA

## Remote interactive pipelines

- Data should be processed close to where it is archived
- Solutions should accommodate both automated pipelines and interactive data processing
- Main advantages of using Jupyter for remote interactive pipelines:
  - **User friendly**: Notebooks are easy and intuitive to use; all results are embedded in a single document
  - **Interactivity is optional**: Notebooks can optionally be run as a non-interactive pipeline for batch processing
  - **Accountability**: Data reduction process is self-documenting and fully repeatable

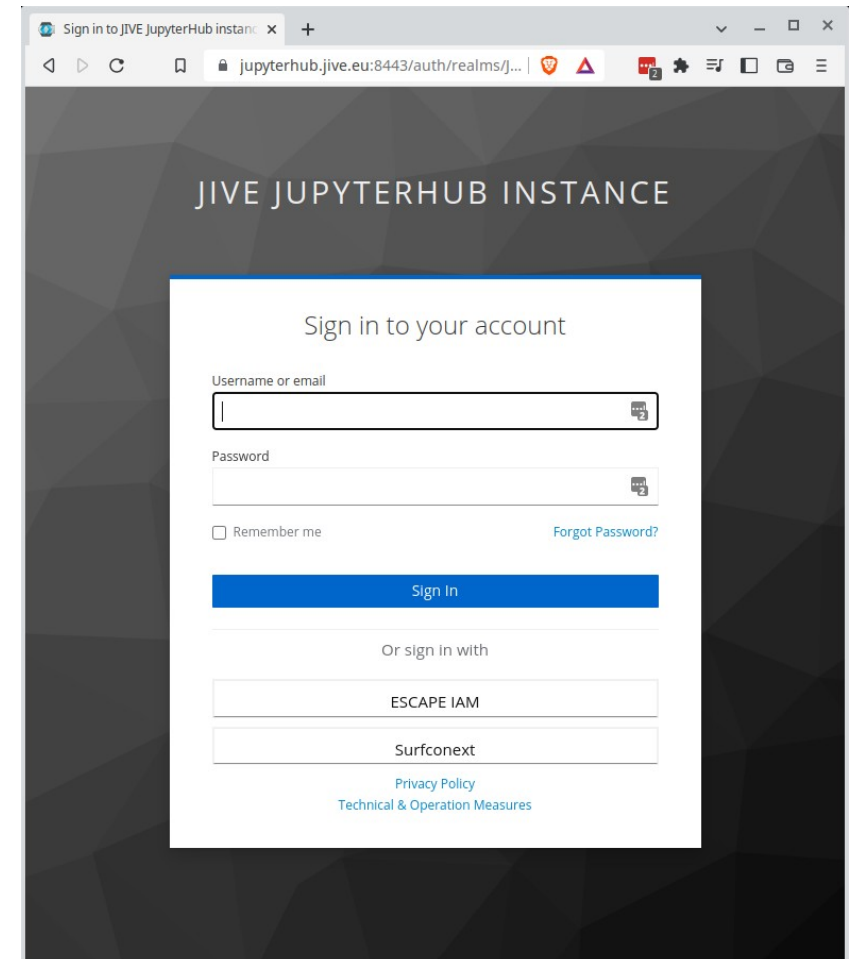
## The Docker image contains

- CASA 6.5.1
- The Jupyter-CASA kernel
- CASA VLBI tools
- EVN data discovery JupyterLab plugin
- A collection of widely used radio astronomy packages



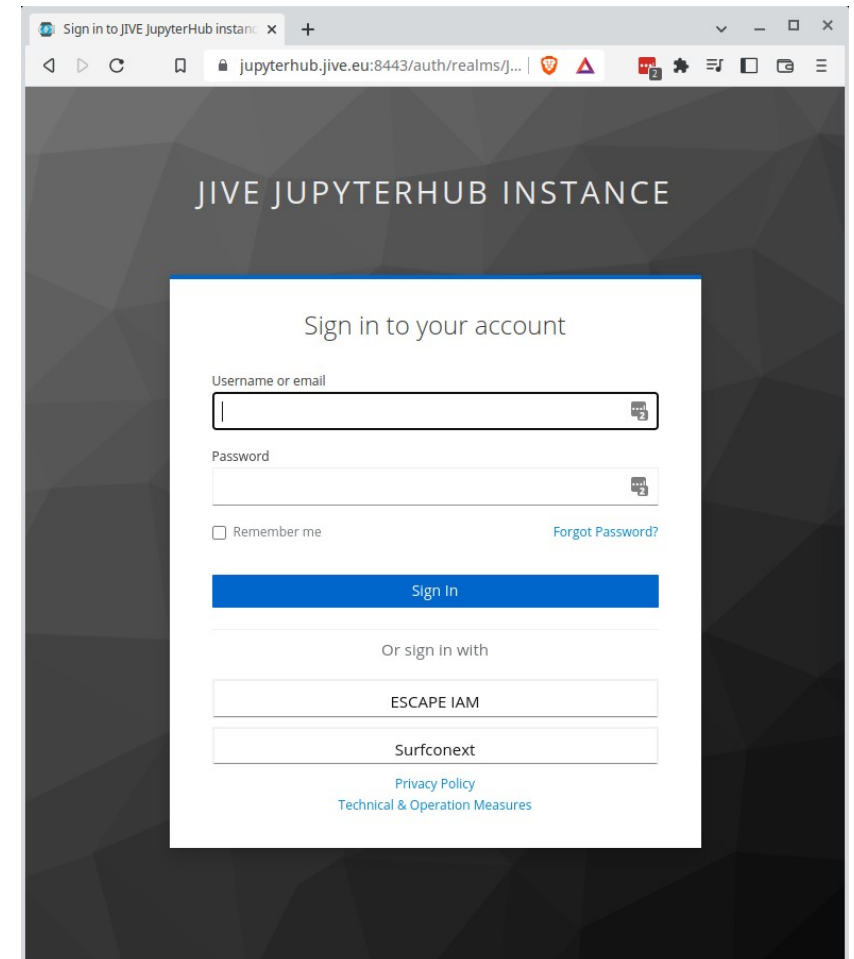
<https://github.com/aardk/jupyter-casa>

- JupyterHub instance hosted at JIVE
- Open service: All interested users can make a free account and process any observation from the EVN archive
- Authentication based on KeyCloak
- Allow federated logins: ESCAPE IAM, SurfCONEXT; Applying for eduGAIN access
- Integrated with the ESAP using BinderHub



<https://jupyterhub.jive.eu>

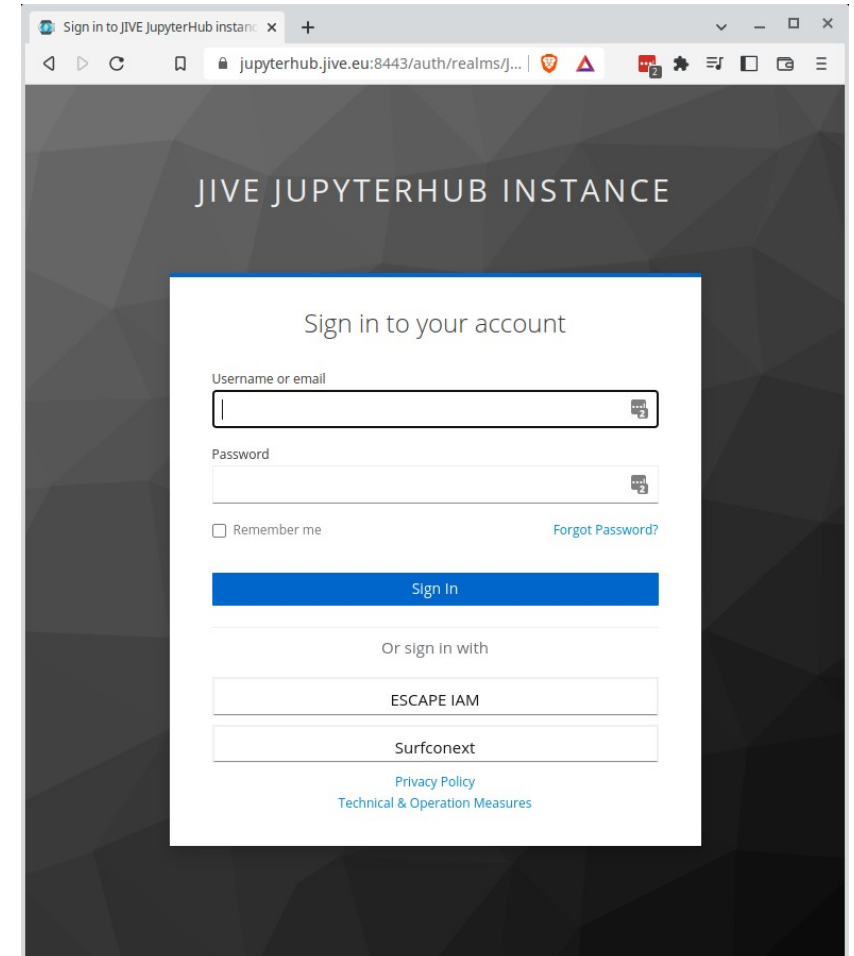
- JupyterHub doesn't handle multiple login methods very easily
- Keycloak supports: Social login, generic OAUTH2, SAML, 2-factor authentication, ...
- Open source software
- Keycloak instance can be shared by multiple services
- Needs external proxy (e.g. SATOSA) for federated logins through eduGAIN



<https://jupyterhub.jive.eu>

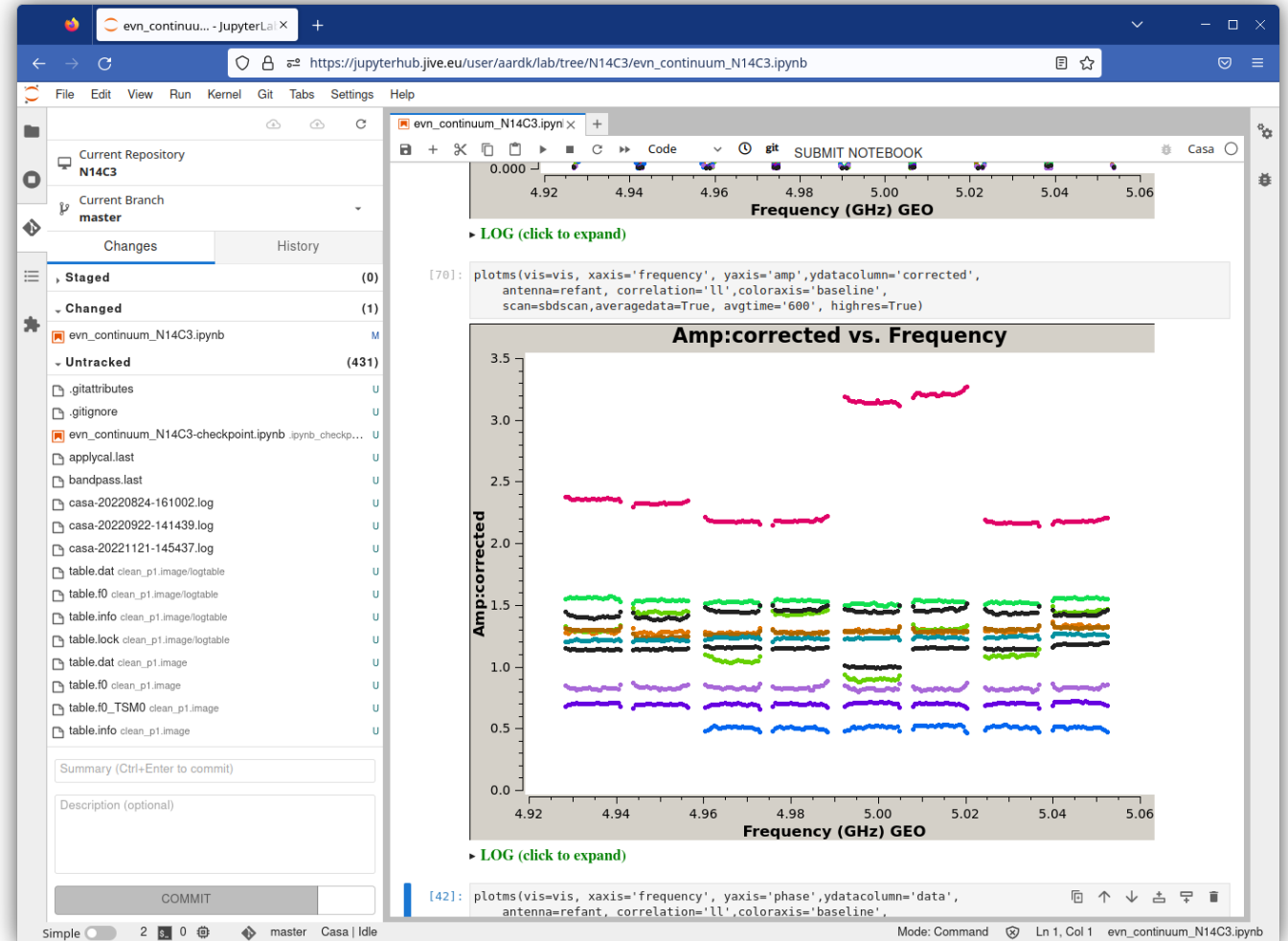


- Data on the EVN archive is
  - **Findable:** Through VO and web interface
  - **Accessible:** EVN data is public after one year and is accessible through the VO or web interface.
  - **Interoperable:** Only open source programs and open data formats are used (FITS, Measurement Set)
  - **Reusable:** Data comes with all relevant metadata and there are no restrictions on use
- Users can submit their notebooks back to the EVN archive allowing other researchers to reproduce their results



<https://jupyterhub.jive.eu>

- Git plugin for JupyterLab
- Adds Git tab view repository
- Easy moving between versions / branches
- Supports syncing to and from remote repositories



The screenshot shows a JupyterLab window with a Git plugin interface on the left and a plot on the right.

**Git Plugin Interface (Left):**

- Current Repository: N14C3
- Current Branch: master
- Changes: (0)
- History: (1)
- Staged: (0)
- Changed: (1)
- Untracked: (431)
- Files listed in untracked: .gitattributes, .gitignore, evn\_continuum\_N14C3-checkpoint.ipynb, applycal.last, bandpass.last, casa-20220824-161002.log, casa-20220922-141439.log, casa-20221121-145437.log, table.dat clean\_p1.image/logtable, table.f0 clean\_p1.image/logtable, table.info clean\_p1.image/logtable, table.lock clean\_p1.image/logtable, table.dat clean\_p1.image, table.f0 clean\_p1.image, table.f0\_TSM0 clean\_p1.image, table.info clean\_p1.image
- Summary (Ctrl+Enter to commit)
- Description (optional)
- COMMIT button

**Plot (Right):**

Frequency (GHz) GEO

LOG (click to expand)

```
[70]: plotms(vis=vis, xaxis='frequency', yaxis='amp', ydatacolumn='corrected',
           antenna=refant, correlation='ll', coloraxis='baseline',
           scan=sbdsan, averagedata=True, avgttime='600', hires=True)
```

**Amp:corrected vs. Frequency**

Y-axis: Amp:corrected (0.0 to 3.5)

X-axis: Frequency (GHz) GEO (4.92 to 5.06)

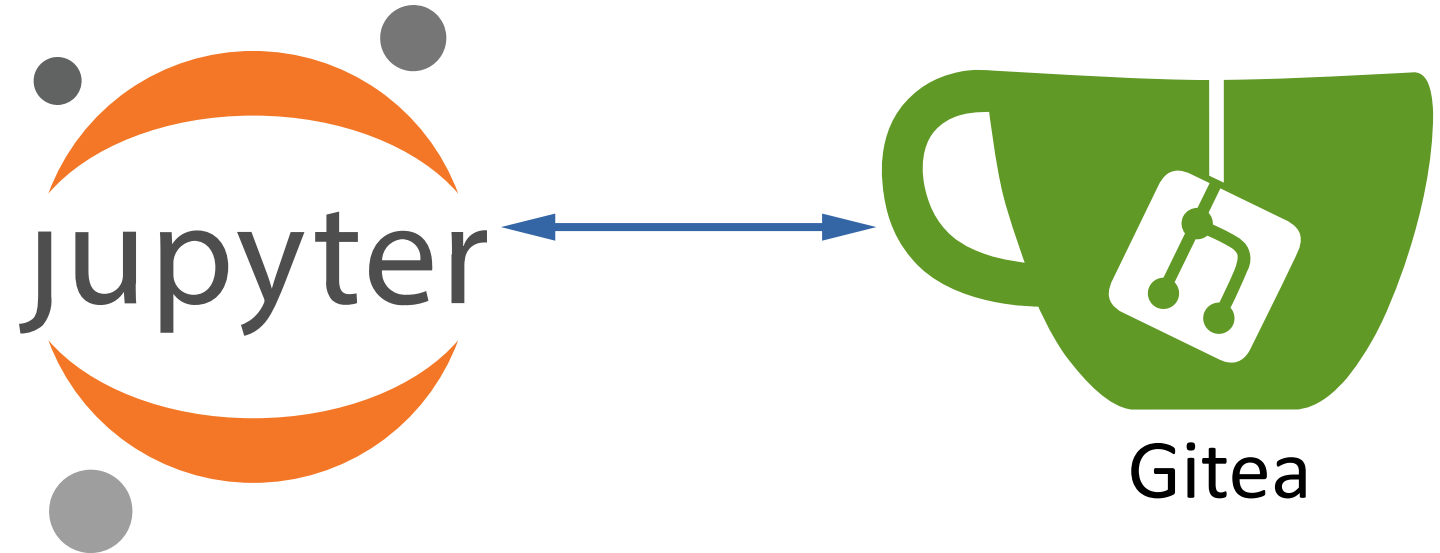
LOG (click to expand)

```
[42]: plotms(vis=vis, xaxis='frequency', yaxis='phase', ydatacolumn='data',
           antenna=refant, correlation='ll', coloraxis='baseline',
```

- Powerful graphical diff
- Shows changed files in working copy
- Similar view to visualize modifications introduced by commits

## User submitted notebooks

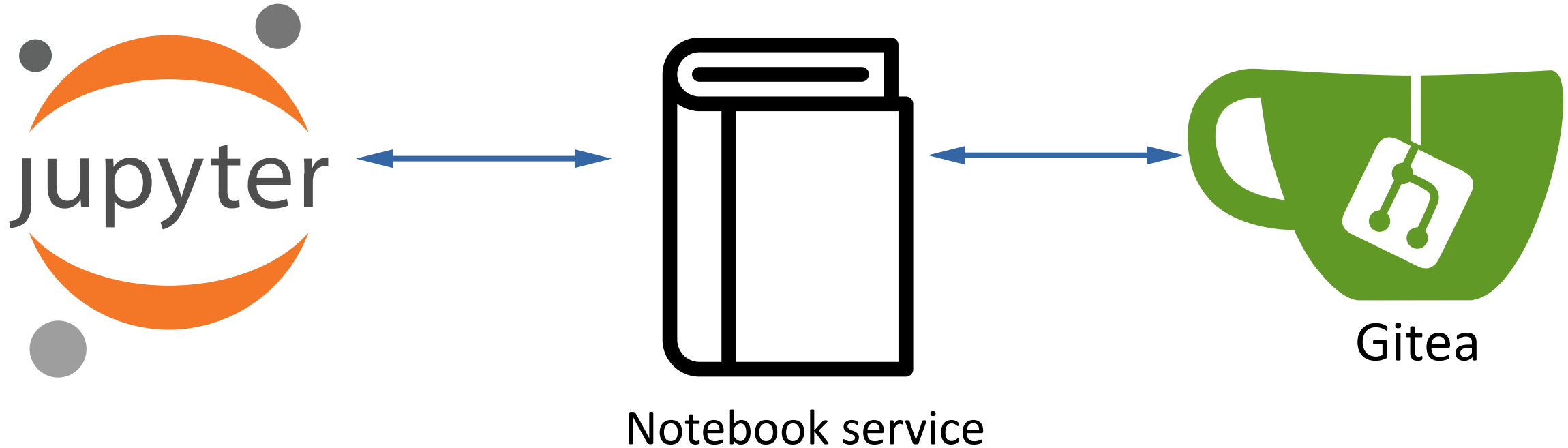
- Why not sync notebooks via git with EVN archive?
- Could be easily implemented using self-hosted service such as Gitea



- Why not sync notebooks via git with EVN archive?
- Could be easily implemented using self-hosted service such as Gitea
- Many users are unwilling to share Git history, only final product



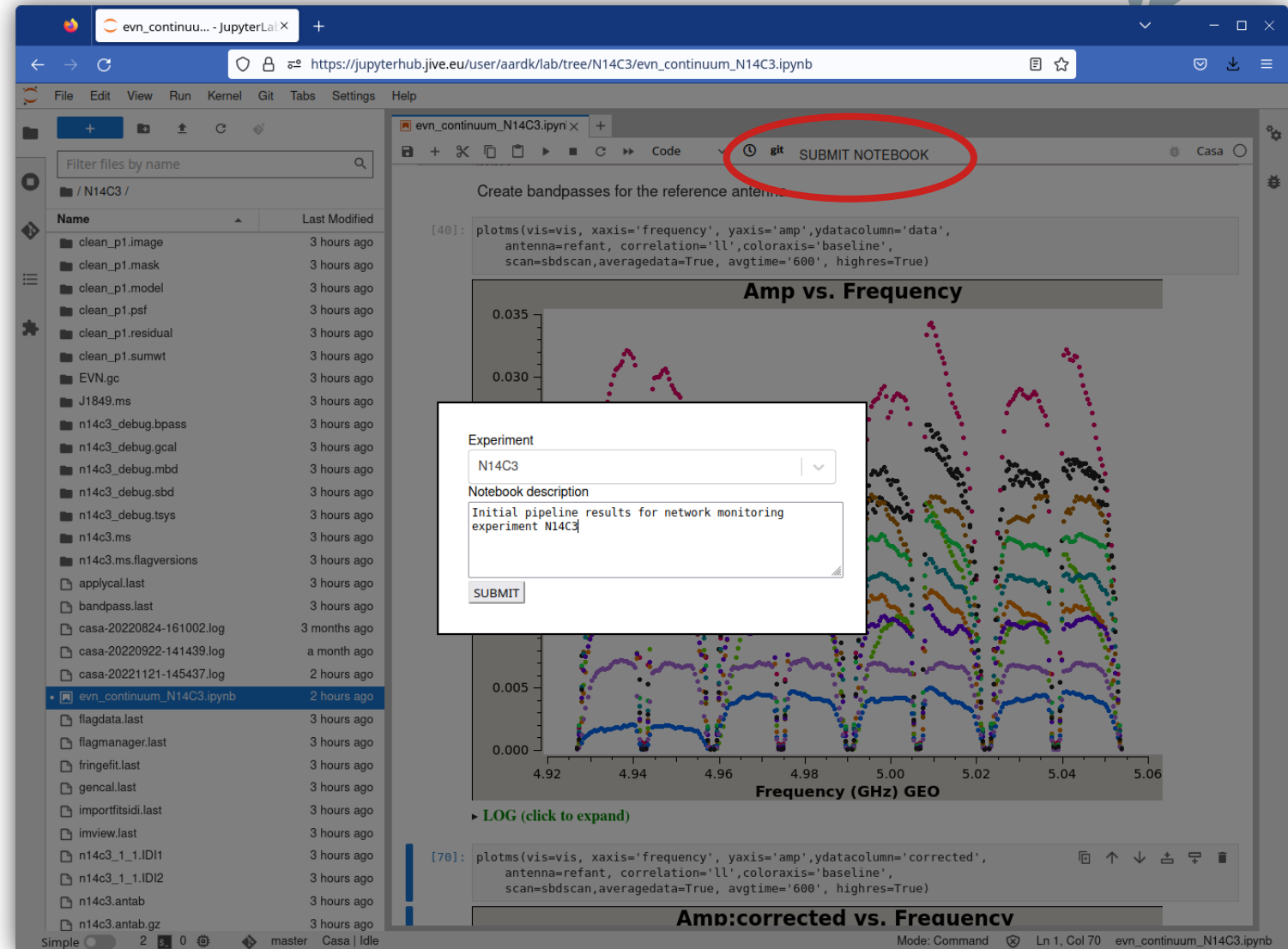
## User submitted notebooks



Notebook service is simple python (Tornado) service that serves as intermediary between JupyterLab and Gitea via a REST API



- We have implemented a JupyterLab plugin to handle notebook submission
- Authentication with notebook service is through OAUTH access tokens
- OAUTH refresh tokens are cached on JupyterLab server so user doesn't have to re-authenticate



The screenshot shows a JupyterLab browser window at `https://jupyterhub.jive.eu/user/aardk/lab/tree/N14C3/evn_continuum_N14C3.ipynb`. The interface includes a file browser on the left, a code editor in the center, and a plot titled "Amp vs. Frequency". A red circle highlights the "SUBMIT NOTEBOOK" button in the top right of the code editor. A modal dialog box is open in the foreground, containing the following information:

```

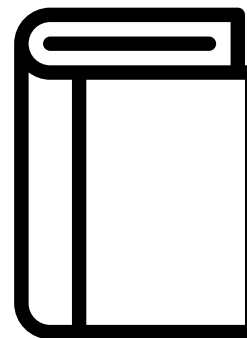
Experiment
N14C3
Notebook description
Initial pipeline results for network monitoring
experiment N14C3
SUBMIT
  
```

The plot shows "Amp vs. Frequency" with the x-axis labeled "Frequency (GHz) GEO" ranging from 4.92 to 5.06 and the y-axis ranging from 0.000 to 0.035. Below the plot, there is a "LOG (click to expand)" section showing code for plotting "Amp:corrected vs. Frequency".

# Authentication Flow



Gitea



Notebook service



Keycloak

Access token

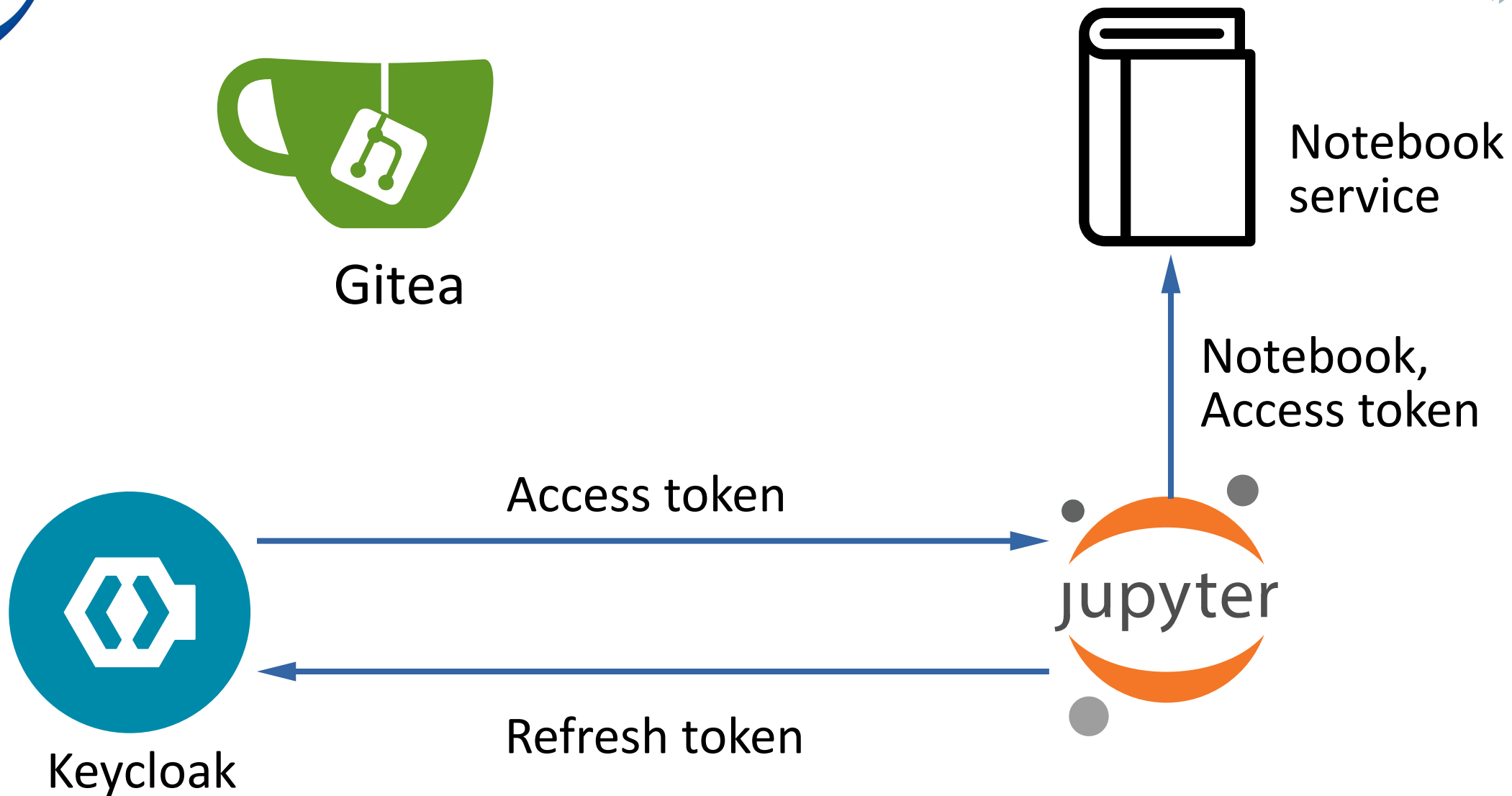


Refresh token





# Authentication Flow



# Authentication Flow



Gitea

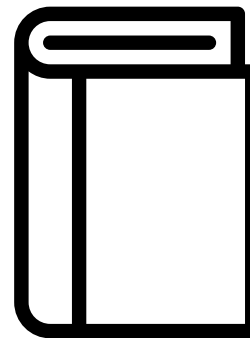


Keycloak



Access token

Refresh token



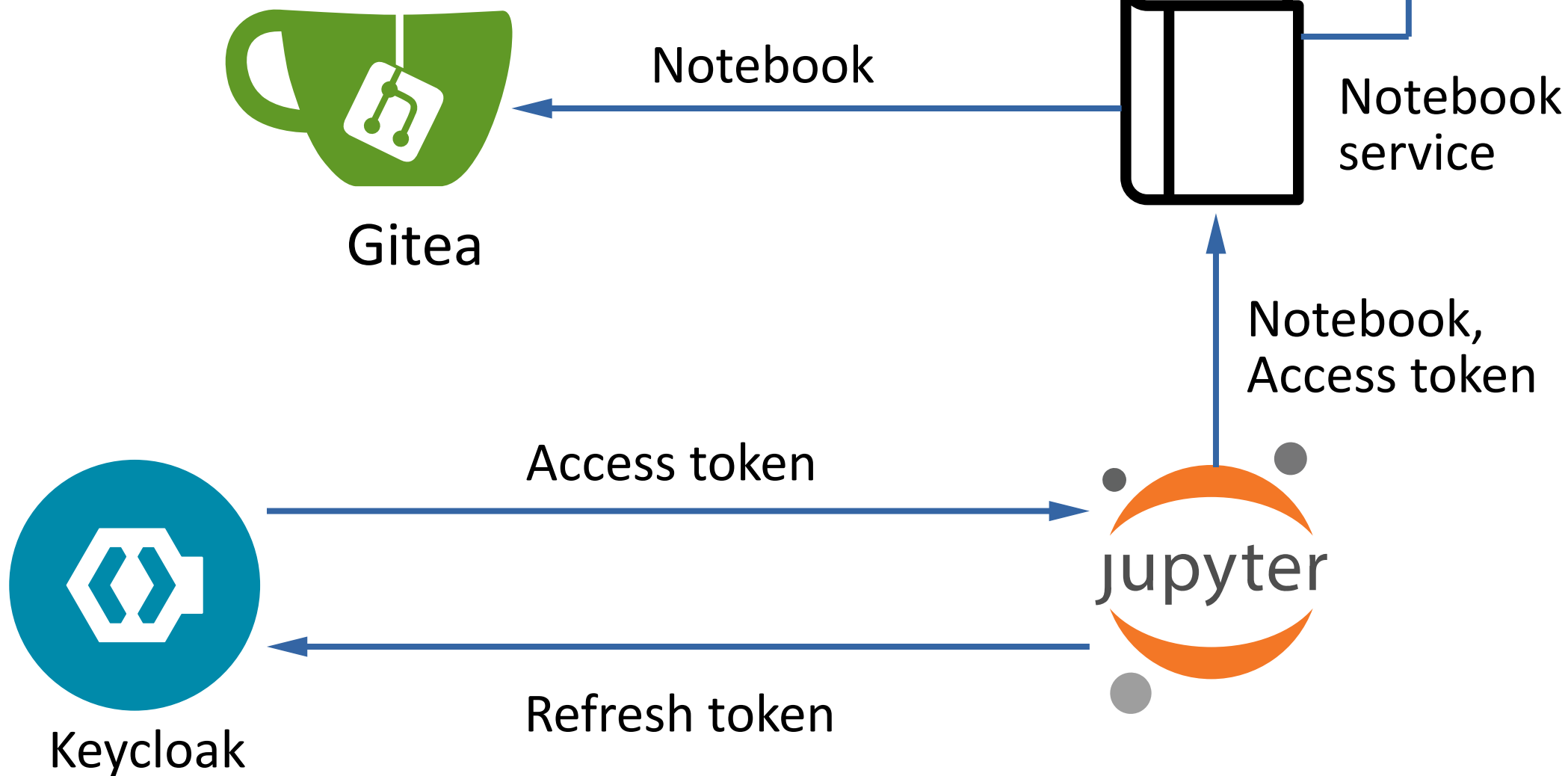
Notebook service

Check token



Notebook,  
Access token

# Authentication Flow



Thank you for your attention!

<https://jupyterhub.jive.eu>