FLATIRON INSTITUTE

Data-sharing at the CCA:

Binder and FlatHUB

SHY GENEL
IVOA May 2021 Interoperability Meeting
Theory/GWS Science Platform Workshop

Team:

Dylan Simon
Elizabeth (Liz) Lovero

Rachel Somerville
Chris Hayward
Shy Genel

binder

binder.flatironinstitute.org

Our goal:

- Provide external users with rudimentary computing power to explore and interact with large data generated or hosted at the Flatiron Institute

Based on BinderHub (mybinder.org), the FI Binder is a Jupyter-based service that allows users to share reproducible interactive computing environments from code repositories

- Allows users to define and provide access to *ephemeral* Jupyter environments

FLATIRON INSTITUTE

binder

binder.flatironinstitute.org

Customizations for Flatiron Cluster:

- Environments can be defined by directories rather than Github repos
- Environments can include datasets hosted locally on cluster
    - Direct, read-only access to multi-PB storage
- Access can be restricted to specific users (via oauth) or open to the public
- Configurable computation limits (cores, memory)

- Currently hosted on-premise: 48-core, 1.5TB memory (adding new host soon at SDSC)
- 10-20 daily users

FLATIRON INSTITUTE

binder

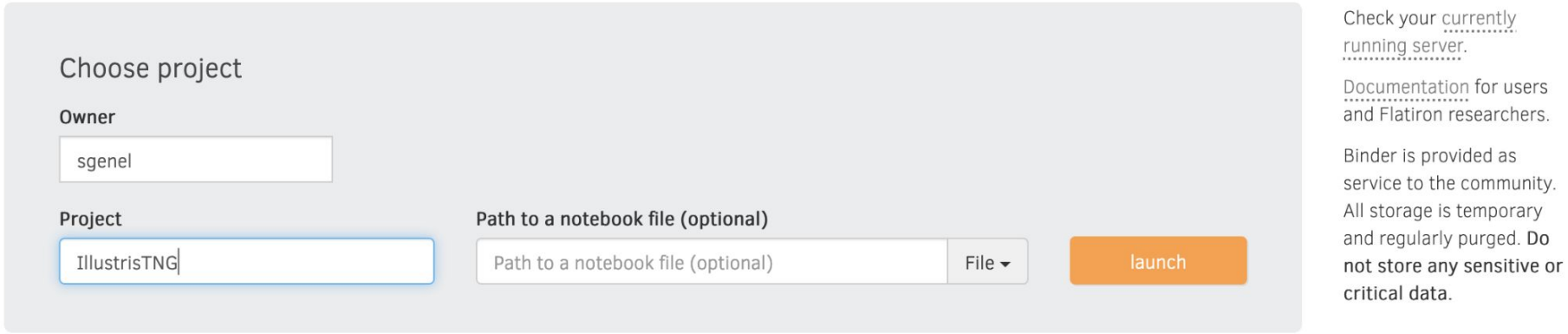
binder.flatironinstitute.org

Environment definition and customization on the FI cluster:

```
sgenel@rusty2:~/public_binder/IllustrisTNG$ ls -l
total 3
-rw-r--r-- 1 sgenel sgenel 285 Jan 29 16:30 environment.yml
lrwxrwxrwx 1 sgenel sgenel  65 Jan 28 17:21 illustris_python -> /mnt/ceph/users/sgenel/BinderUsers/IllustrisTNG/illustris_python/
lrwxrwxrwx 1 sgenel sgenel  69 Feb 18 15:34 illustris_sam -> /mnt/ceph/users/sgenel/BinderUsers/IllustrisTNG/illustris_sam-master/
-rw-r--r-- 1 sgenel sgenel 166 Sep 20  2019 postBuild
lrwxrwxrwx 1 sgenel sgenel  57 Sep 20  2019 SimulationData -> /mnt/ceph/users/sgenel/PUBLIC/Illustris_IllustrisTNG_all/
sgenel@rusty2:~/public_binder/IllustrisTNG$ cat environment.yml
dependencies:
  - python=3.7.3
  - numpy
  - pip
  - vim
  - emacs
  - jupyter_contrib_nbextensions
  - gsl
  - swig
  - pip:
    - nbgitpuller
    - sphinx-gallery
    - pandas
    - matplotlib
    - astropy
    - matplotlib
sgenel@rusty2:~/public_binder/Mack10$ cat .public_binder
users:
- sgenel@flatironinstitute.org
cpu_guarantee: 10
cpu_limit: 10
mem_limit: '20G'
```

FLATIRON INSTITUTE

Starting the environment from an external user's access point:



FLATIRON INSTITUTE

Logout | Control Panel

Files | Running | Clusters | Nbextensions

Select items to perform actions on them.

Upload | New |

0 / SimulationData

Name	Last Modified	File size
..	seconds ago	
Illustris-1	a year ago	
Illustris-3	a year ago	
IllustrisTNG100	5 months ago	
IllustrisTNG100-2	5 months ago	
IllustrisTNG300	5 months ago	
IllustrisTNG300-2	5 months ago	
L25n128TNG	6 months ago	
L25n256TNG	6 months ago	
L25n512TNG	6 months ago	
L75n1820TNG_DM	5 months ago	
L75n455TNG	5 months ago	
TNG100-1	2 months ago	
TNG100-2	a year ago	
TNG100-3	a year ago	

FlatHUB

flathub.flatironinstitute.org

Flatiron Institute Data Exploration and Comparison Hub (FlatHUB) is a public data exploration query interface to tabular, numeric data (primarily catalogs).

- Built on Elasticsearch, allows efficient filtering, summary statistics, aggregate data (histograms)
- Provides an interactive web interface, HTTP API, Python client
- Currently running on 8-node cluster
    - 5B rows, 400B indexed values, across 11 catalogs, 2TB indexed data on 64TB NVMe

Performance

- GAIA DR2 (1.7B rows, 92 columns) = 10~20 seconds/query (uncached)
- Smaller, cached, and filtered datasets can be much faster
- Bulk downloads of raw data with many fields can be slow (due mainly to ES JSON interface)

FLATIRON INSTITUTE

Flatiron Institute Data Exploration and Comparison Hub

Flatiron Institute Data Exploration and Comparison Hub (FlatHUB) is a science platform for diverse types of data, that allows users to explore and compare data from different simulations and datasets with one another and with curated observational/experimental collections. Users can browse and filter the data collections, make simple preview plots, and download sub-samples of the data.

Catalogs

Ananke: GAIA on FIRE

CANDELS

SC-SAM

UniverseMachine

Lu-SAM

GAEA

Gaia DR2

Collections

Milky Way

Cosmological Galaxy Formation Simulations

Cosmological Dark Matter-Only Simulations

CANDELS

Illustris Subfind Subhalos

PLOT TYPE:

scatterplot

X-AXIS:

VelDisp

LIN LOG

Y-AXIS:

BHMass

LIN LOG

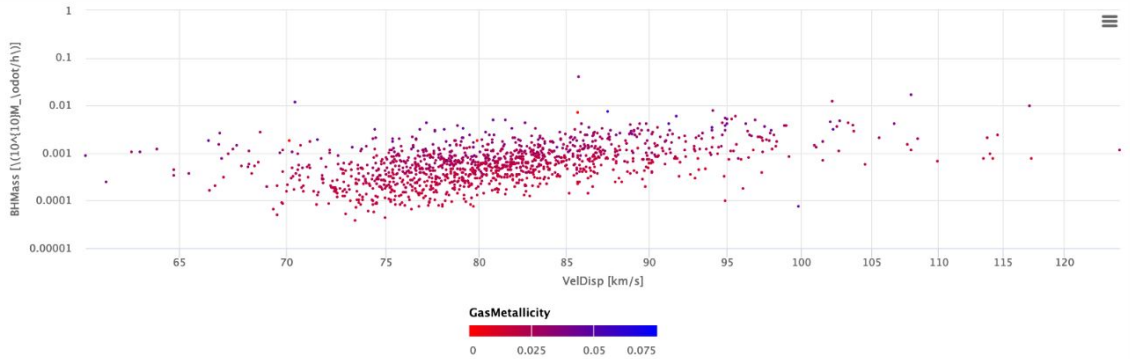
Z-AXIS:

Choose Z-Axis...

COLOR:

GasMetallicity

Click and drag to zoom into the figure. Filters in the right column reflect plot selections.



Filter Python All Fields About

Active Filters RESET ALL

simulation Illustris-1 (40385)

snap 135 135 μ: 106.08728488300112 Range: 54 - 135

BHMass \(10^{10}M_\odot/h\) RESET

0.0000194 0.3195247

μ: 0.000013812435 Range: 0.0000000 - 1.4466752

MassInRadType stars \(10^{10}M_\odot/h\) RESET

1 1.5

μ: 0.84027101 Range: 0.00046021744 - 150.08513

VelDisp km/s RESET

60.913688 124.69671

μ: 81.495308 Range: 60.913689 - 124.69672

Select field to filter MassInRadType stars

VIEW RAW DATA 1,273 RESULTS (FILTERED FROM 1,175,372,132) ACTIVE FILTERS simulation snapshot BHMass MassInRadType_star VelDisp FORMAT Choose format... DOWNLOAD

Illustris Subfind Subhalos

PLOT TYPE:

scatterplot

X-AXIS:

VelDisp

LIN LOG

Y-AXIS:

BHMass

LIN LOG

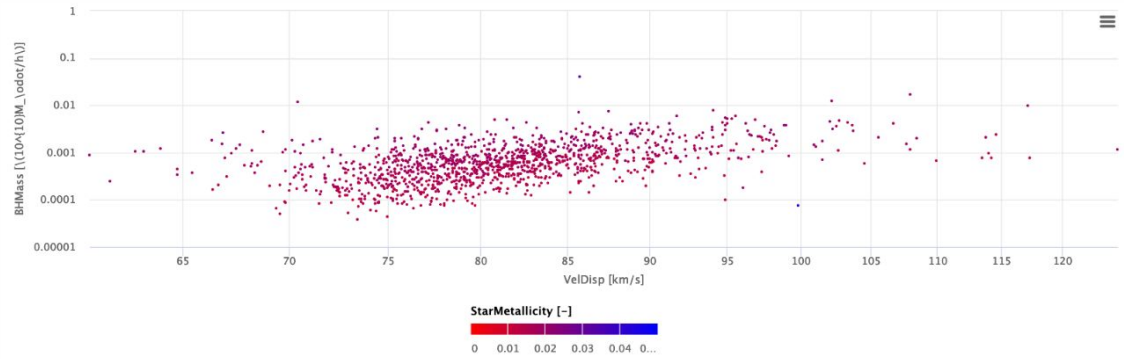
Z-AXIS:

Choose Z-Axis...

COLOR:

StarMetallicity

Click and drag to zoom into the figure. Filters in the right column reflect plot selections.



Filter Python All Fields About

Python Query

Example python code to apply the above filters and retrieve data. To use, download and install this module .

```
import flathub.client
illustris_sub = flathub.client.Catalog("illustris_sub",
q = illustris_sub.query(fields = ["simulation","redshif
  simulation = "0",
  snapshot = 135,
  BHMass = (0.00019421746814890246, 0.3195247505759213
  MassInRadType_stars = (1, 1.5),
  VelDisp = (60.91368865966797, 124.69671630859375))
dat = q.numpy()
```

VIEW RAW DATA 1,273 RESULTS (FILTERED FROM 1,175,372,132) ACTIVE FILTERS simulation snapshot BHMass MassInRadType_star VelDisp FORMAT

Choose format...
raw data
csv
csv.gz
ecsv
ecsv.gz
npy
npy.gz
attachment files

DOWNLOAD

Coming up: the *Compare* feature

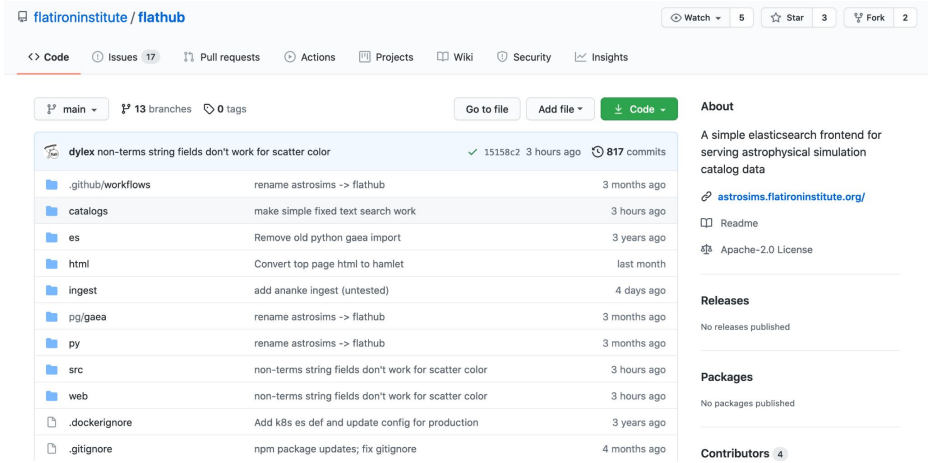
- Integrate additional cosmological simulations: IllustrisTNG, EAGLE, Bahamas, SIMBA, FIRE, …
- Challenges:
    - Comparing like-for-like quantities
    - Keeping documentation (as well as data) up to date

In the works:

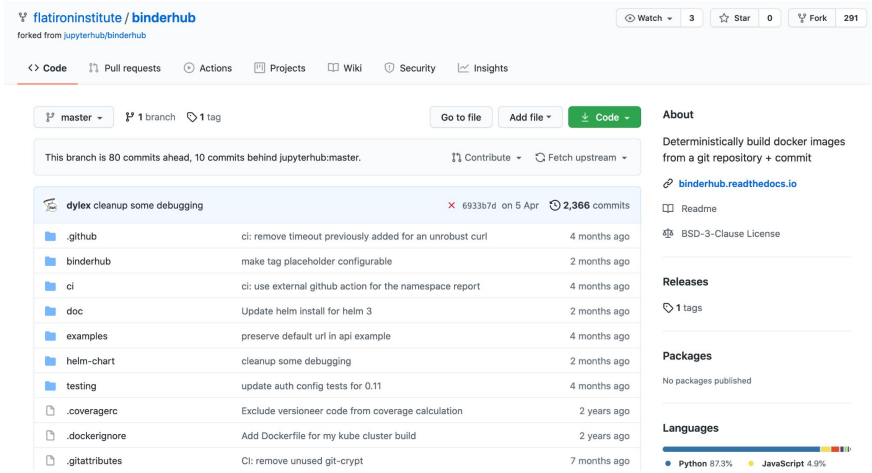
- How to serve (large) particle data: connect to external resources? Host mirror FlatHUB servers where the data is?

Both FlatHUB and the FI Binder are open source

https://github.com/flatironinstitute/flathub



https://github.com/flatironinstitute/binderhub



FLATIRON INSTITUTE


FLATIRON
INSTITUTE