

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

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

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

Environment definition and customization on the FI cluster:

```
sgene1@rusty2:~/public_binder/IllustrisTNG$ ls -l
total 3
-rw-r--r-- 1 sgene1 sgene1 285 Jan 29 16:30 environment.yml
lrwxrwxrwx 1 sgene1 sgene1 65 Jan 28 17:21 illustris_python -> /mnt/ceph/users/sgene1/BinderUsers/IllustrisTNG/illustris_python/
lrwxrwxrwx 1 sgene1 sgene1 69 Feb 18 15:34 illustris_sam -> /mnt/ceph/users/sgene1/BinderUsers/IllustrisTNG/illustris_sam-master/
-rw-r--r-- 1 sgene1 sgene1 166 Sep 20 2019 postBuild
lrwxrwxrwx 1 sgene1 sgene1 57 Sep 20 2019 SimulationData -> /mnt/ceph/users/sgene1/PUBLIC/Illustris_IllustrisTNG_all/
sgene1@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

sgene1@rusty2:~/public_binder/Mack10$ cat .public_binder
users:
- sgene1@flatironinstitute.org
cpu_guarantee: 10
cpu_limit: 10
mem_limit: '20G'
```

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



Choose project

Owner

sgenel

Project

IllustrisTNG

Path to a notebook file (optional)

Path to a notebook file (optional)

File ▾

launch

Check your currently running server.

Documentation for users and Flatiron researchers.

Binder is provided as service to the community. All storage is temporary and regularly purged. **Do not store any sensitive or critical data.**

[Files](#)
[Running](#)
[Clusters](#)
[Nbextensions](#)

Select items to perform actions on them.

[Download](#)
[Upload](#)
[New](#)
[Refresh](#)

<input type="checkbox"/> 0 ▼ / SimulationData		Name ▼	Last Modified	File size
<input type="checkbox"/> 📁 ..			seconds ago	
<input type="checkbox"/>	📁 Illustris-1		a year ago	
<input type="checkbox"/>	📁 Illustris-3		a year ago	
<input type="checkbox"/>	📁 IllustrisTNG100		5 months ago	
<input type="checkbox"/>	📁 IllustrisTNG100-2		5 months ago	
<input type="checkbox"/>	📁 IllustrisTNG300		5 months ago	
<input type="checkbox"/>	📁 IllustrisTNG300-2		5 months ago	
<input type="checkbox"/>	📁 L25n128TNG		6 months ago	
<input type="checkbox"/>	📁 L25n256TNG		6 months ago	
<input type="checkbox"/>	📁 L25n512TNG		6 months ago	
<input type="checkbox"/>	📁 L75n1820TNG_DM		5 months ago	
<input type="checkbox"/>	📁 L75n455TNG		5 months ago	
<input type="checkbox"/>	📁 TNG100-1		2 months ago	
<input type="checkbox"/>	📁 TNG100-2		a year ago	
<input type="checkbox"/>	📁 TNG100-3		a year ago	

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 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

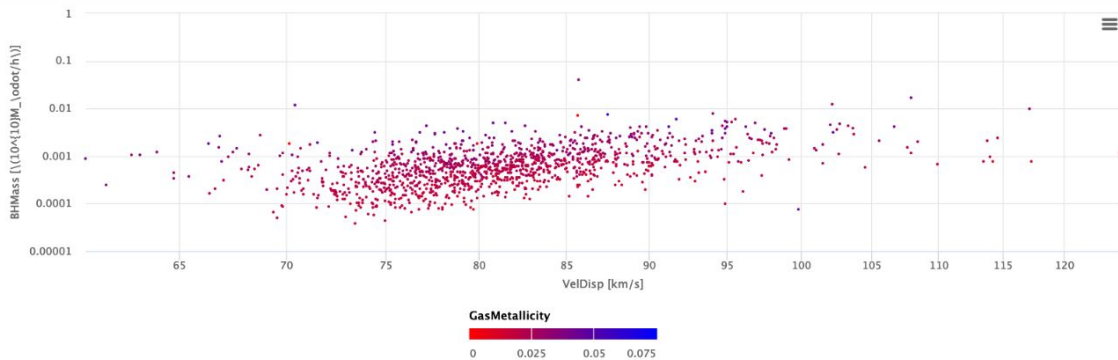
Z-AXIS:

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

 μ :

106.08728488300112

Range: 54

135

135

✕

BHMass $\backslash(10^4)(10M_{\odot}/h)$

RESET

0.0000194

0.3195247

 μ : 0.000013812435

Range: 0.0000000 - 1.4466752

✕

MassInRadType stars $\backslash(10^4)(10M_{\odot}/h)$

RESET

1

1.5

 μ : 0.84027101

Range: 0.00046021744 - 150.08513

✕

VelDisp km/s

RESET

60.91368E

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:

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.

StarMetallicity [-]
0 0.01 0.02 0.03 0.04 0...

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","redshift
simulation = "0",
snapshot = 135,
BHMass = (0.000019421746814890246, 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... **DOWNLOAD**

- raw data
- csv
- csv.gz
- ecsv
- ecsv.gz
- npz
- npz.gz
- attachment files

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>

The screenshot shows the GitHub repository page for `flatironinstitute/flathub`. The repository is on the `main` branch and has 13 branches and 0 tags. It has 5 watchers, 3 stars, and 2 forks. The repository description is "A simple elasticsearch frontend for serving astrophysical simulation catalog data". The repository is licensed under Apache-2.0. The repository has 817 commits and 17 issues. The repository is forked from `flatironinstitute/flathub`. The repository is licensed under Apache-2.0. The repository has 817 commits and 17 issues. The repository is forked from `flatironinstitute/flathub`.

File	Commit Message	Time
<code>.github/workflows</code>	rename astrosims -> flathub	3 months ago
<code>catalogs</code>	make simple fixed text search work	3 hours ago
<code>es</code>	Remove old python gaea import	3 years ago
<code>html</code>	Convert top page html to hamlet	last month
<code>ingest</code>	add ananke ingest (untested)	4 days ago
<code>pg/gaea</code>	rename astrosims -> flathub	3 months ago
<code>py</code>	rename astrosims -> flathub	3 months ago
<code>src</code>	non-terms string fields don't work for scatter color	3 hours ago
<code>web</code>	non-terms string fields don't work for scatter color	3 hours ago
<code>.dockerignore</code>	Add kBs es def and update config for production	3 years ago
<code>.gitignore</code>	npm package updates; fix gitignore	4 months ago

<https://github.com/flatironinstitute/binderhub>

The screenshot shows the GitHub repository page for `flatironinstitute/binderhub`. The repository is on the `master` branch and has 1 branch and 1 tag. It has 3 watchers, 0 stars, and 291 forks. The repository description is "Deterministically build docker images from a git repository + commit". The repository is licensed under BSD-3-Clause License. The repository has 2,366 commits and 17 issues. The repository is forked from `flatironinstitute/binderhub`. The repository is licensed under BSD-3-Clause License. The repository has 2,366 commits and 17 issues. The repository is forked from `flatironinstitute/binderhub`.

File	Commit Message	Time
<code>.github</code>	ci: remove timeout previously added for an unrobust curl	4 months ago
<code>binderhub</code>	make tag placeholder configurable	2 months ago
<code>ci</code>	ci: use external github action for the namespace report	4 months ago
<code>doc</code>	Update helm install for helm 3	2 months ago
<code>examples</code>	preserve default url in api example	4 months ago
<code>helm-chart</code>	cleanup some debugging	2 months ago
<code>testing</code>	update auth config tests for 0.11	4 months ago
<code>.coveragerc</code>	Exclude versioneer code from coverage calculation	2 years ago
<code>.dockerignore</code>	Add Dockerfile for my kube cluster build	2 years ago
<code>.gitattributes</code>	ci: remove unused git-crypt	7 months ago

