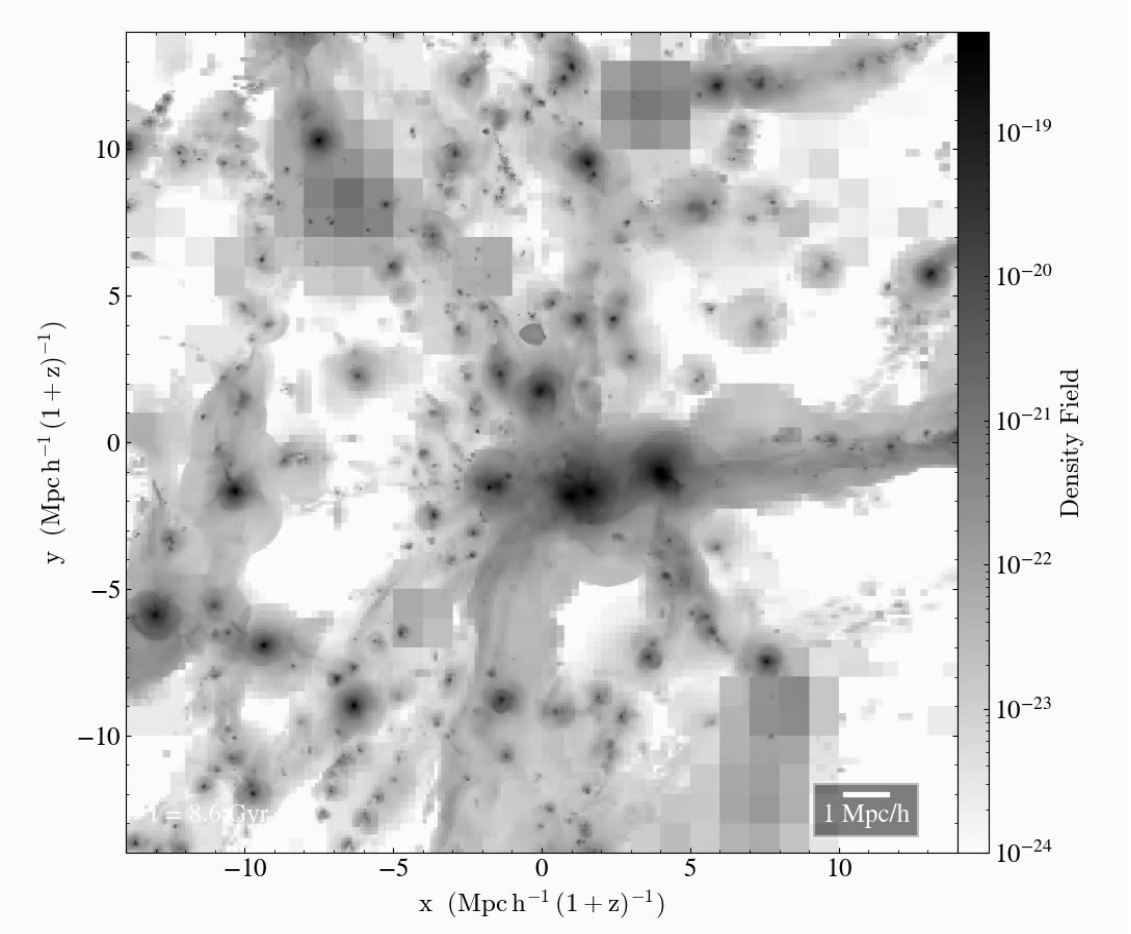
Use Case of reana in Astrophysics

Prateek Gupta

Research Associate

Matthias Hoeft, Harry Enke, Arman Khalatyan, Manuel Giffels, Benoit Roland, Christoph Wissing, Oliver Freyermuth, Thomas Schoerner-Sadenius, Elena Sacchi, Olaf Michaelis, Philip Bechtle, Michael Hübner, Alexander Drabent

Cosmological Simulation



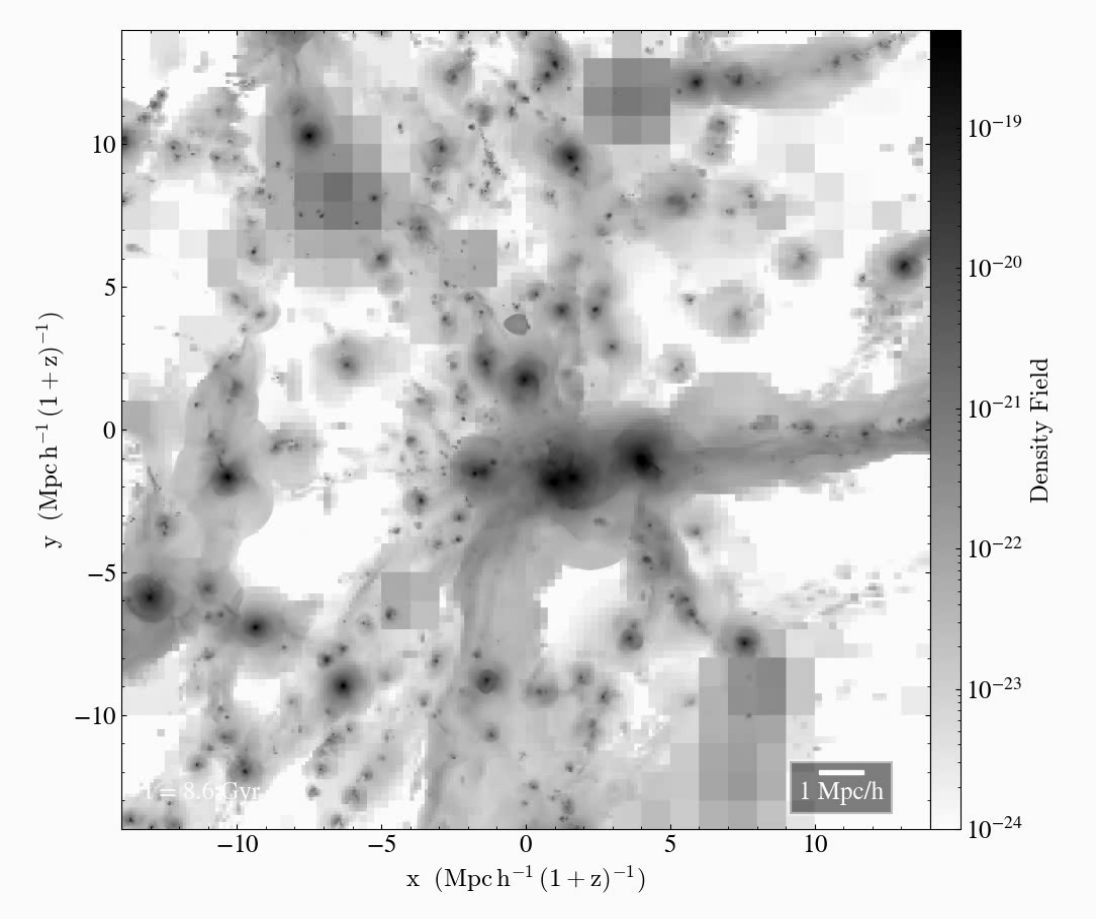
- Simulate the evolution of the structures of the Universe.

To perform these simulations, huge computational resources required:

> 1000 CPU
> 20000h CPU
> 1 TB RAM

- Demands computational resources

Cosmological Simulation



Typical cosmological simulations occupies ~500 GB to 10 TB of disk size.

- Demands large storage

Post-processing of such large volumetric data:

- Projection plot
- Volume rendering
- Time-series
- Finding galaxy groups/clusters
- Implementing numerical models

- Demands computational resources

Support for different compute backend

Why

reana

???

Allow to open interactive session on same workflow

Containerise once, reuse elsewhere

Multiple environment in a single workflow

Support different Computational workflow engines

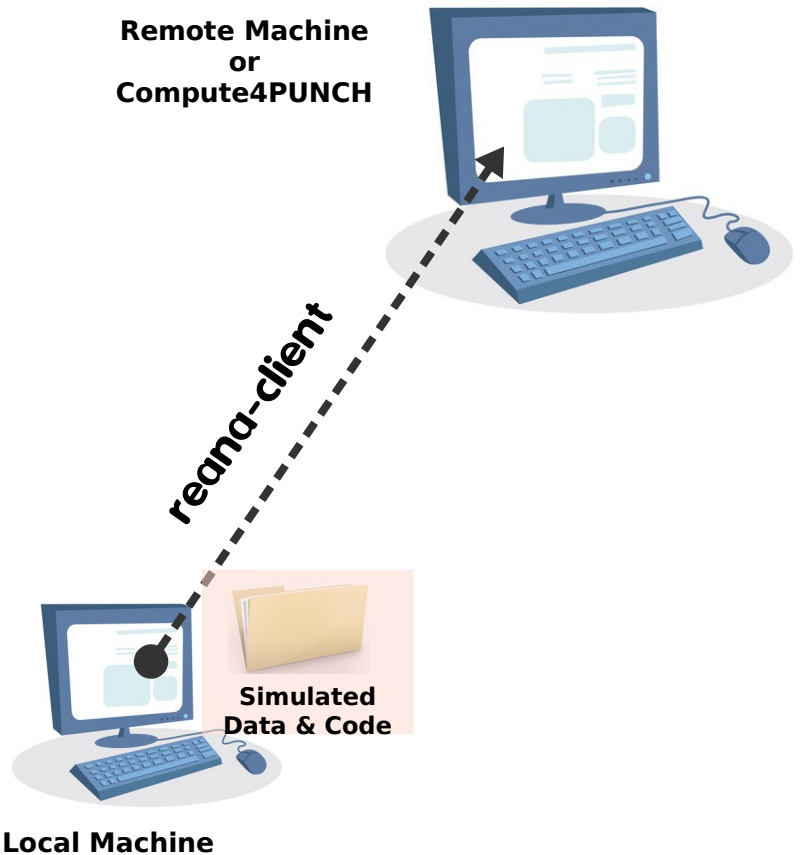
How does reana make this simpler?

```yaml
inputs:
  directories:
    - Simulated_Data
  files:
    - multiple_plot_test.py
  parameters:
    maincode: multiple_plot_test.py
```

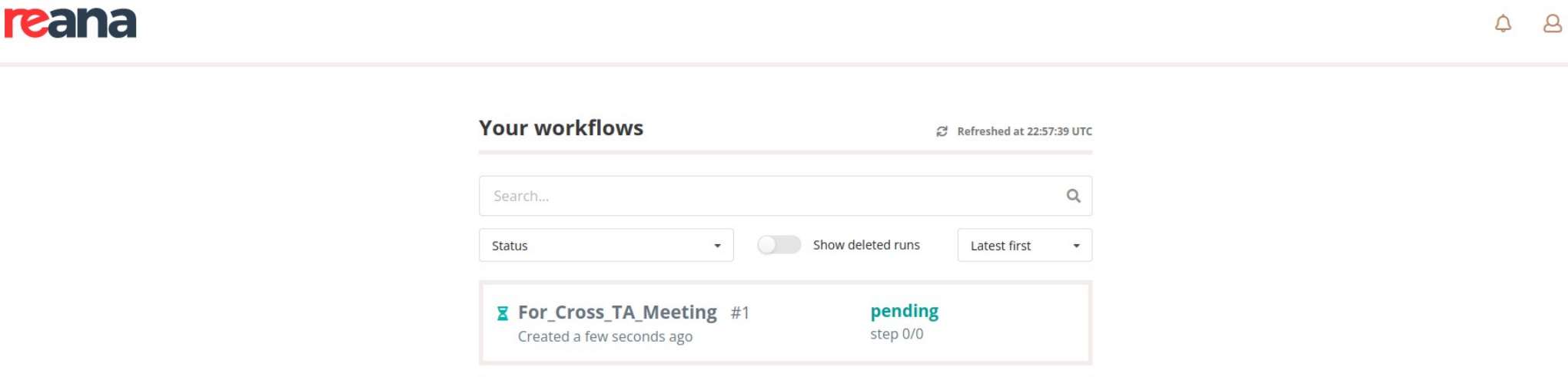
```yaml
workflow:
  type: serial
  specification:
    steps:
      - environment: 'docker.io/tlsprateek/ytproject'
        commands:
          - mkdir Output
          - python3 "${maincode}"
```

```yaml
outputs:
  directories:
    - Output
```

reana.yaml

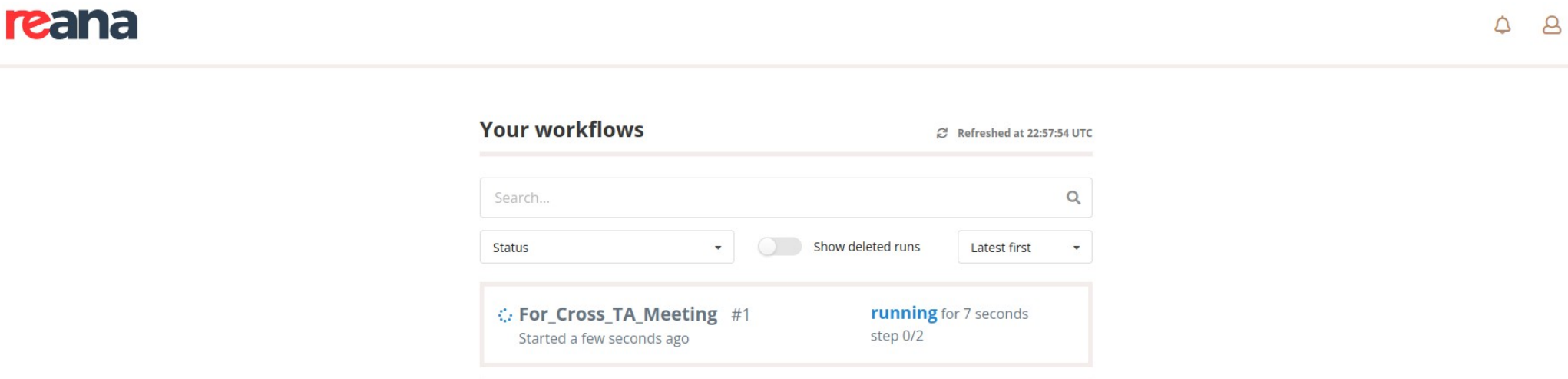


reana Graphical web-interface



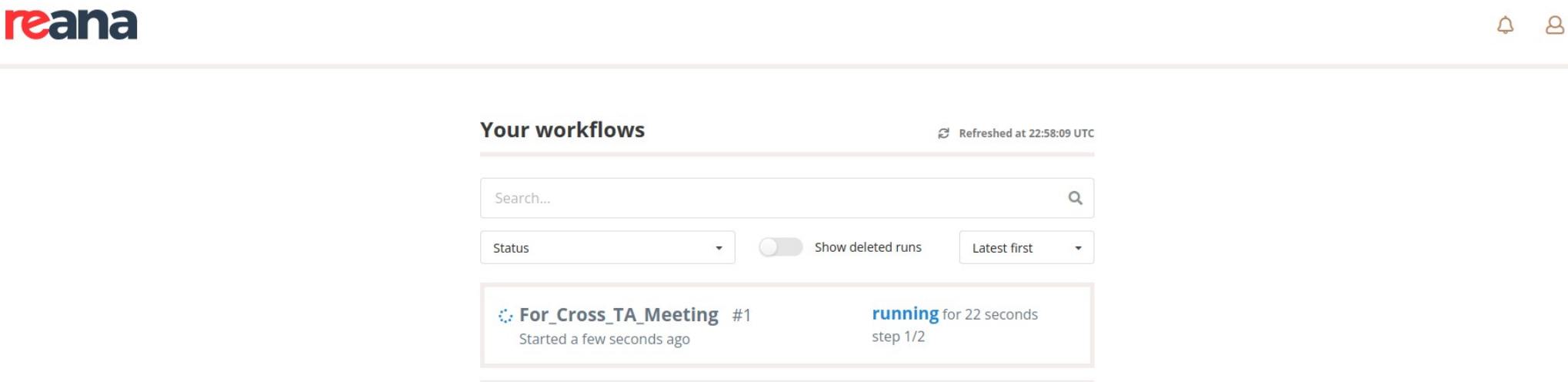
Workflow namespace : For_Cross_TA_Meeting

reana Graphical web-interface



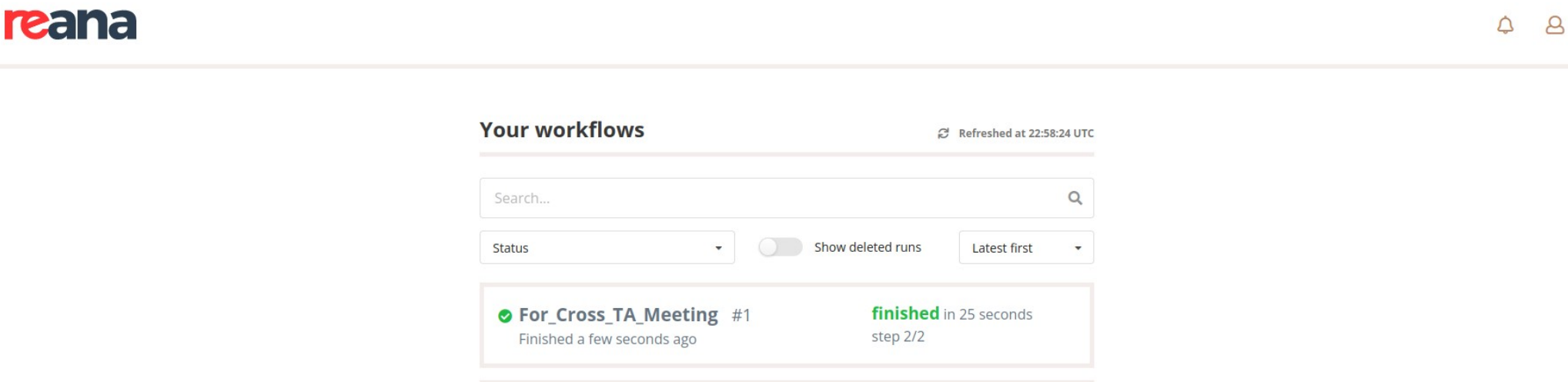
Workflow namespace : For_Cross_TA_Meeting

reana Graphical web-interface



Workflow namespace : For_Cross_TA_Meeting

reana Graphical web-interface



Workflow namespace : For_Cross_TA_Meeting

reana

For_Cross_TA_Meeting #1
Finished a few seconds ago

finished in 25 seconds
step 2/2

Engine logs | Job logs | Workspace | Specification

Step reana-run-job-175cb3d9-119d-40aa-... | finished in 15 seconds | Kubernetes | docker.io/tlsprateek/ytproject | $ python3 "multiple_plot_test.py"

```
yt : [INFO     ] 2024-06-11 22:58:01,623 Parameters: current_time              = 302.05401068366 code_time
yt : [INFO     ] 2024-06-11 22:58:01,623 Parameters: domain_dimensions         = [256 256 256]
yt : [INFO     ] 2024-06-11 22:58:01,623 Parameters: domain_left_edge          = [0.47070312 0.46289062 0.46044922] code_length
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: domain_right_edge         = [0.53320312 0.52539062 0.52294922] code_length
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: cosmological_simulation   = 1
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: current_redshift          = -2.2204460492503e-16
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: omega_lambda              = 0.7257
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: omega_matter              = 0.2743
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: omega_radiation           = 0.0
yt : [INFO     ] 2024-06-11 22:58:01,624 Parameters: hubble_constant           = 0.702
yt : [INFO     ] 2024-06-11 22:58:02,829 xlim = 0.470703 0.533203
yt : [INFO     ] 2024-06-11 22:58:02,829 ylim = 0.462891 0.525391
yt : [INFO     ] 2024-06-11 22:58:02,831 xlim = 0.470703 0.533203
yt : [INFO     ] 2024-06-11 22:58:02,831 ylim = 0.462891 0.525391
yt : [INFO     ] 2024-06-11 22:58:02,835 Making a fixed resolution buffer of (('grid', 'y-velocity')) 800 by 800
yt : [INFO     ] 2024-06-11 22:58:03,329 Making a fixed resolution buffer of (('grid', 'Density')) 800 by 800
yt : [INFO     ] 2024-06-11 22:58:03,406 Making a fixed resolution buffer of (('grid', 'x-velocity')) 800 by 800
yt : [INFO     ] 2024-06-11 22:58:03,547 Making a fixed resolution buffer of (('grid', 'z-velocity')) 800 by 800


Completed
```

reana

For_Cross_TA_Meeting #1
Finished 2 minutes ago

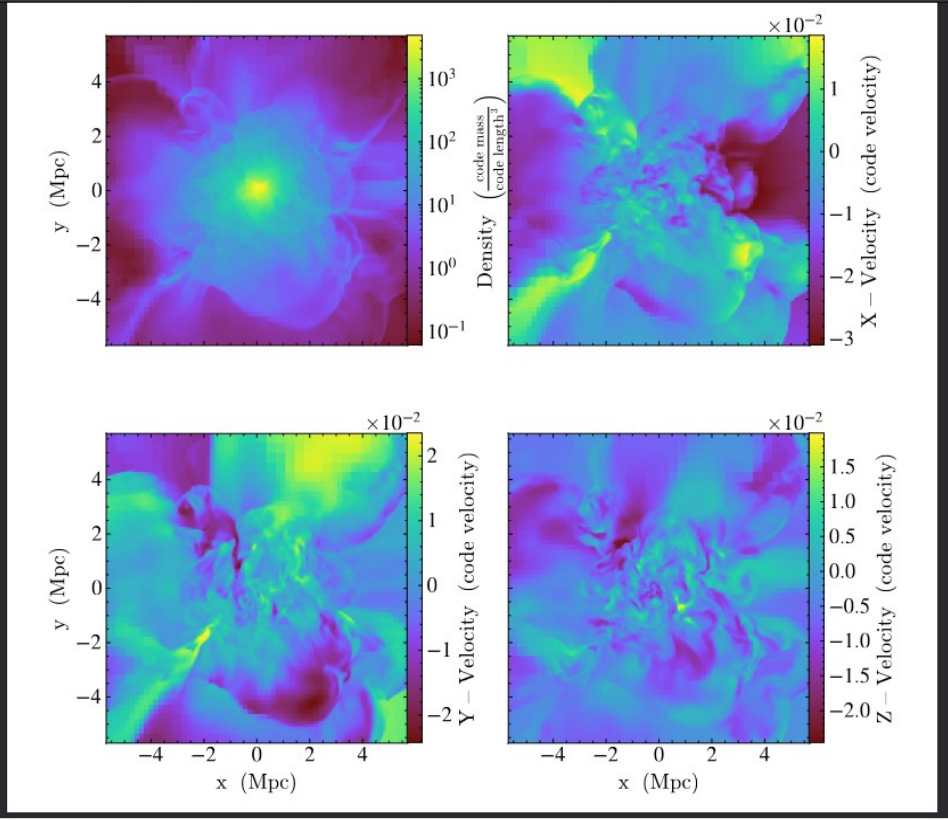
finished in 25 seconds
step 2/2

Engine logs | Job logs | **Workspace** | Specification

Search...

Name	Modified	Size
multiple_plot_test.py	2024-06-11T22:57:24	582 Bytes
reana.yaml	2024-06-11T22:57:24	379 Bytes
Output/multiplot_export_to_mpl.pdf	2024-06-11T22:58:06	232.94 KiB
Simulated_Data/simdata.h5	2024-06-11T22:57:36	3 GiB

Output/multiplot_export_to_mpl.pdf



Download

What if a slight modification is required in the same workflow?

Allow to open interactive session on same workflow

For_Cross_TA_Meeting #1
Finished 25 minutes ago

finished in 25 seconds
step 2/2

Open Jupyter Notebook
Delete workflow

Engine logs | Job logs | **Workspace** | Specification

Search...

Name	Modified	Size
multiple_plot_test.py	2024-06-11T22:57:24	582 Bytes
reana.yaml	2024-06-11T22:57:24	379 Bytes
Output/multiplot_export_to_mpl.pdf	2024-06-11T22:58:06	232.94 KiB
Simulated_Data/simdata.h5	2024-06-11T22:57:36	3 GiB

reana

Success!
The interactive session has been created. However, it could take several minutes to start the Jupyter Notebook. Click on the Jupyter logo to access it.

For_Cross_TA_Meeting #1
Finished 26 minutes ago

finished in 25 seconds
step 2/2

Engine logs | Job logs | **Workspace** | Specification

Search...

Name	Modified	Size
multiple_plot_test.py	2024-06-11T22:57:24	582 Bytes
reana.yaml	2024-06-11T22:57:24	379 Bytes
Output/multiplot_export_to_mpl.pdf	2024-06-11T22:58:06	232.94 KiB
Simulated_Data/simdata.h5	2024-06-11T22:57:36	3 GiB

jupyter

Quit Logout

Files Running Clusters

Select items to perform actions on them.

Upload New

0 /	Name	Last Modified	File size
Output		26 minutes ago	
Simulated_Data		27 minutes ago	
multiple_plot_test.py		27 minutes ago	582 B
reana.yaml		27 minutes ago	379 B

jupyter multiple_plot_test.py 9 minutes ago

Logout

File Edit View Language Python

```python
import yt

ds = yt.load("Simulated_Data/simdata.h5")

# velocity_Turbulence
fields = [
    ("grid", "Density"),
    ("grid", "x-velocity"),
    ("grid", "y-velocity"),
    ("grid", "z-velocity"),
]

p = yt.SlicePlot(ds, "z", fields)
p.set_log(("grid", "x-velocity"), False)
p.set_log(("grid", "y-velocity"), False)
p.set_log(("grid", "z-velocity"), False)

# this returns a matplotlib figure with an ImageGrid and the slices
# added to the grid of axes (in this case, 2x2)
fig = p.export_to_mpl_figure((2, 2))

fig.tight_layout()

fig.savefig("Output/multiplot_export_to_mpl.pdf")
```

jupyter multiple_plot_test.py✔ a few seconds ago

Logout

File Edit View Language Python

```python
import yt

ds = yt.load("Simulated_Data/simdata.h5")

# velocity_Turbulence
fields = [
    ("grid", "Density"),
    ("grid", "x-velocity"),
    ("grid", "y-velocity"),
    ("grid", "Temperature"),
]

p = yt.SlicePlot(ds, "z", fields)
p.set_log(("grid", "x-velocity"), False)
p.set_log(("grid", "y-velocity"), False)
#p.set_log(("grid", "z-velocity"), False)

# this returns a matplotlib figure with an ImageGrid and the slices
# added to the grid of axes (in this case, 2x2)
fig = p.export_to_mpl_figure((2, 2))

fig.tight_layout()

fig.savefig("Output/New_multiplot_export_to_mpl.pdf")
```

reana

Your workflows

Refreshed at 23:26:49 UTC

Search...

Status | Show deleted runs | Latest first

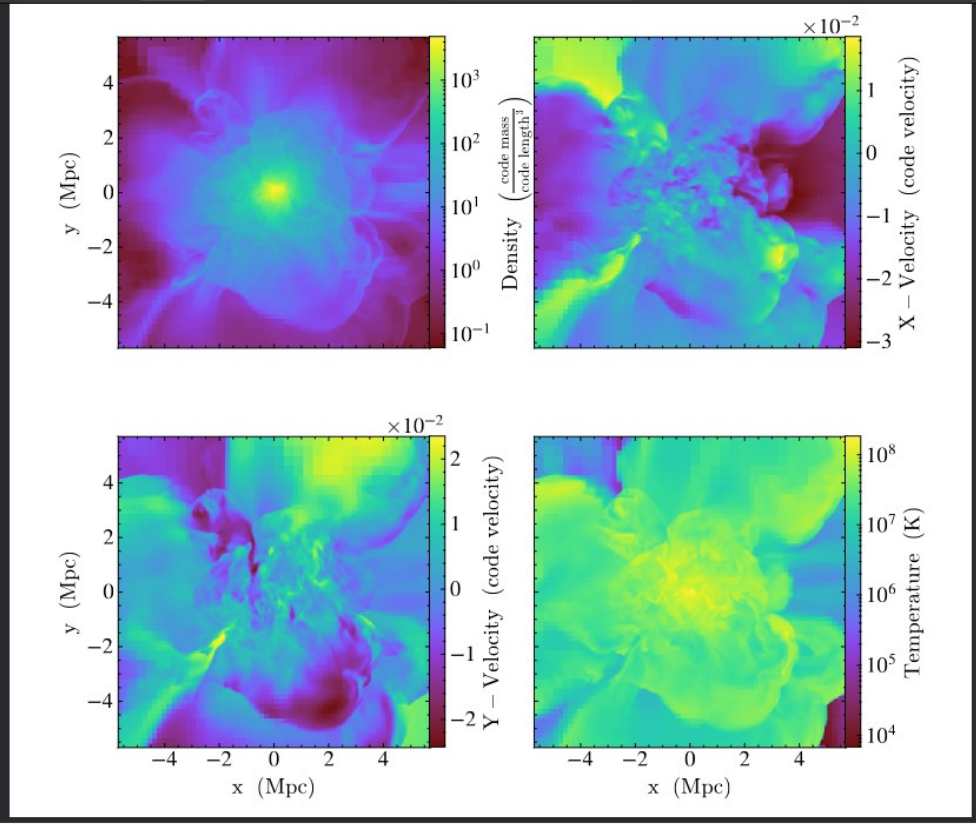
For_Cross_TA_Meeting #1.1
Started a few seconds ago

running for 1 seconds
step 0/2

For_Cross_TA_Meeting #1
Finished 29 minutes ago

finished in 25 seconds
step 2/2

New_multiplot_export_to_mpl.pdf

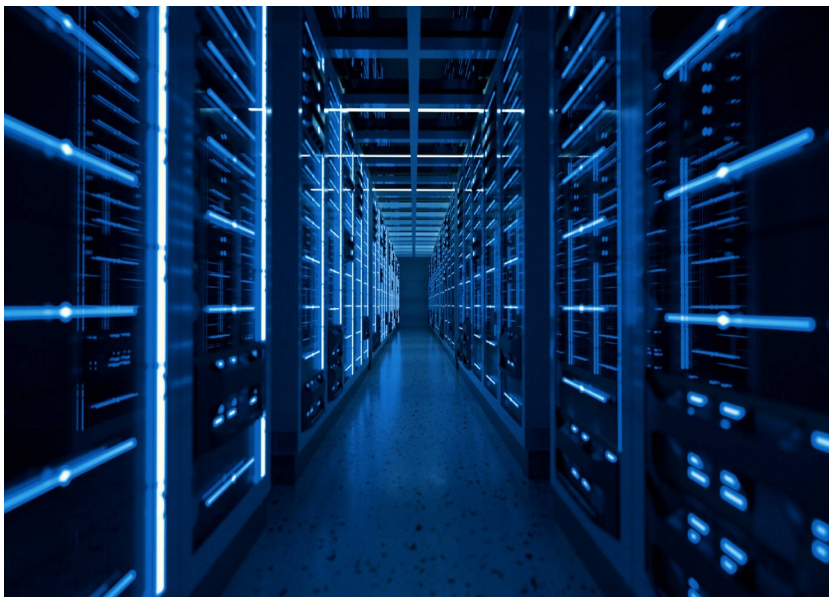


Download

Remote Machine or Compute4PUNCH

Simulated Data

Local Machine

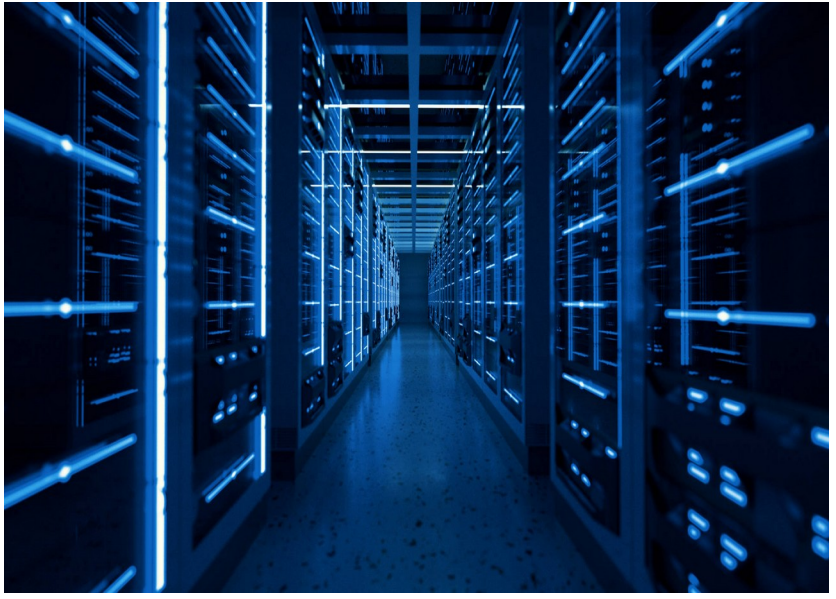


Storage4PUNCH

Remote Machine
or
Compute4PUNCH

Simulated Data

Local Machine



Storage4PUNCH

Remote Machine
or
Compute4PUNCH

Submit and
Run workflow
on Remote
Machine

Local Machine

Storage4PUNCH

Remote Machine
or
Compute4PUNCH

Workflow will query the simulated data from Storage4PUNCH

Local Machine

Storage4PUNCH

How to access Storage4PUNCH ?

With reana !!!

Need :

- **A web-client: DaviX**
- **Access Token**

Caution :

- **How to handle Access Token safely ?**
- **Where to store such a sensitive information ?**

reana offers a secret store:

- secretly add the Access Token

- Either as environment variable or as a secret file.

Oh !!!, Now we need to add DaviX in our old analysis container.

reana !!!

Yes :)

Multiple environment in a single workflow

reana.yaml

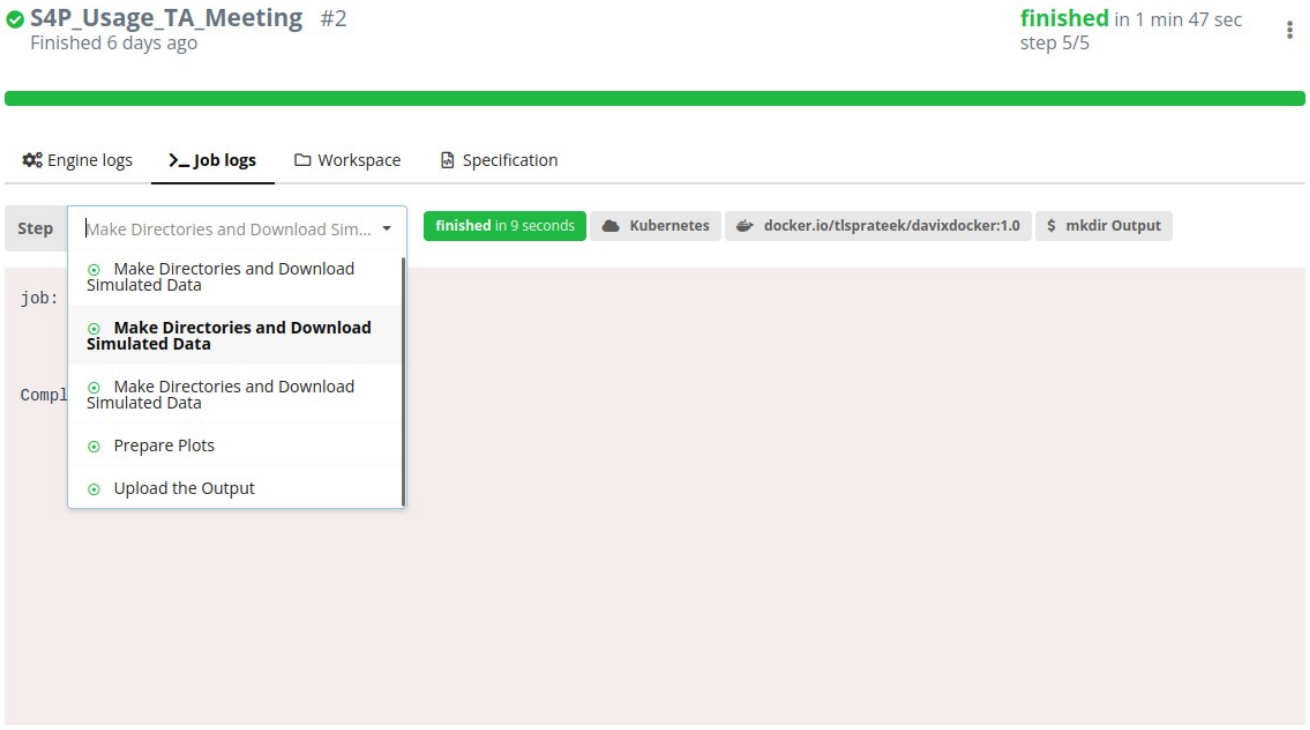
```yaml
inputs:
  files:
    - multiple_plot_test.py
    - download_data.sh
    - upload_data.sh
  parameters:
    maincode: multiple_plot_test.py
    getdata: download_data.sh
    uploadResult: upload_data.sh

workflow:
  type: serial
  specification:
    steps:
      - name: Make Directories and Download Simulated Data
        environment: 'docker.io/tlsprateek/davixdocker:1.0'
        commands:
          - mkdir Simulated_Data
          - mkdir Output
          - bash "${getdata}"

      - name: Prepare Plots
        environment: 'docker.io/tlsprateek/ytproject'
        commands:
          - python3 "${maincode}"

      - name: Upload the Output
        environment: 'docker.io/tlsprateek/davixdocker:1.0'
        commands:
          - bash "${uploadResult}"
```

reana


S4P_Usage_TA_Meeting #2
Finished 6 days ago
finished in 1 min 47 sec
step 5/5
Engine logs
Job logs
Workspace
Specification
Step
Make Directories and Download Sim...
finished in 9 seconds
Kubernetes
docker.io/tlsprateek/davixdocker:1.0
$ mkdir Output
job:
Compl
Make Directories and Download Simulated Data
Make Directories and Download Simulated Data
Make Directories and Download Simulated Data
Prepare Plots
Upload the Output

reana

S4P_Usage_TA_Meeting #2
Finished 6 days ago

finished in 1 min 47 sec
step 5/5

Engine logs | Job logs | **Workspace** | Specification

Search...

Name	Modified	Size
download_data.sh	2024-06-12T01:28:37	180 Bytes
multiple_plot_test.py	2024-06-12T01:28:37	582 Bytes
reana.yaml	2024-06-12T01:28:37	786 Bytes
upload_data.sh	2024-06-12T01:28:38	195 Bytes
Output/multiplot_export_to_mpl.pdf	2024-06-12T01:30:21	232.94 KiB
Simulated_Data/simdata.h5	2024-06-12T01:30:04	3 GiB

Next Steps

Trying to do LoFAR data processing.

We are able to run the LoFAR pipeline With Reana
- CWL Workflow
- Test LoFAR data ~ 20 GB

With Reana – able to submit workflow on Compute4PUNCH machine.

Currently, testing the processing of large data ~ 4 TB

```
inputs:
  directories:
    - LINC
    - Data_all
  files:
    - concat_prefactor_all.json
  parameters:
    input: concat_prefactor_all.json
workflow:
  type: serial
  specification:
    steps:
      - environment: 'linc-wn:latest'
        compute_backend: compute4punch
        commands:
          - cwltool --preserve-entire-environment --no-container
            --parallel --log-dir "LINC/log_store"
            LINC/workflows/HBA_calibrator.cwl ${input}
```

Take-home messages

reana is a powerful tool.

- Submit jobs/workflow on remote servers.
- Use Containers: a big relief from installing softwares on every machine.
- One can use multiple containers in a single workflow.
- Allows to open an interactive session in the same workspace in one click; using Jupyter-Notebook

Thank you