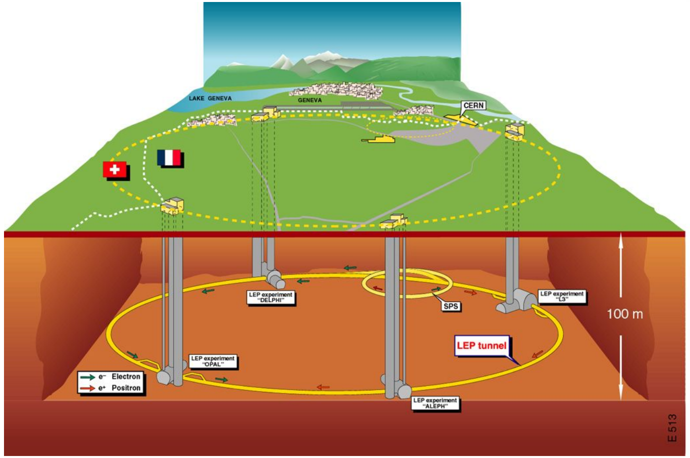
Open and Reproducible Research Services in LHC Particle Physics

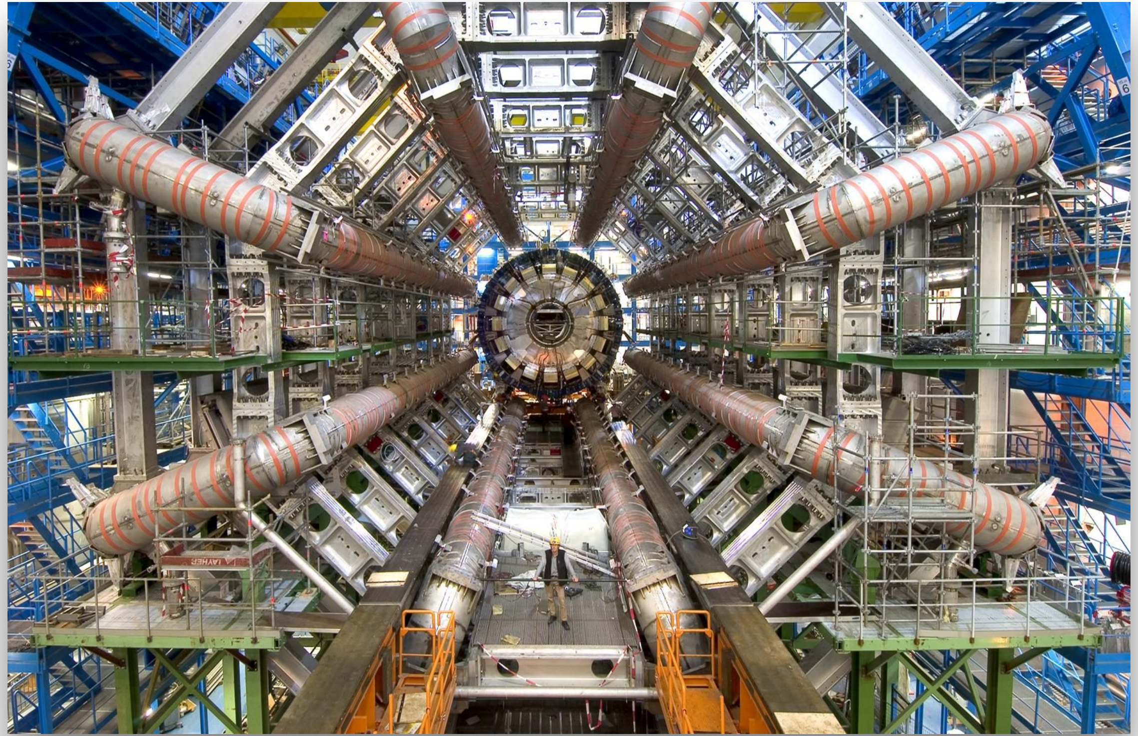
Diego Rodríguez
CERN

Webinar: Toward Open Reproducible Research, 2021-02-10, reusableresearch.com

CERN Large Hadron Collider

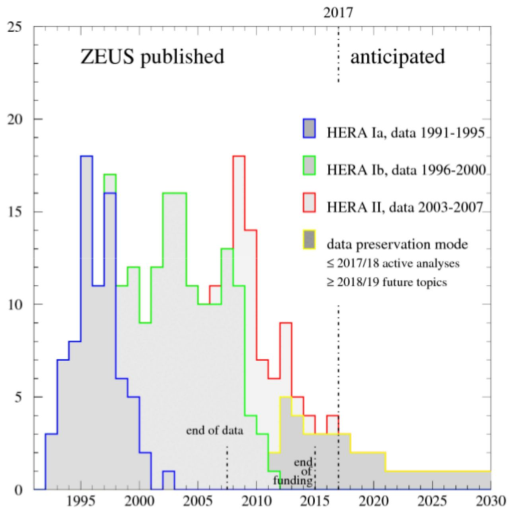


http://cds.cern.ch/record/842153

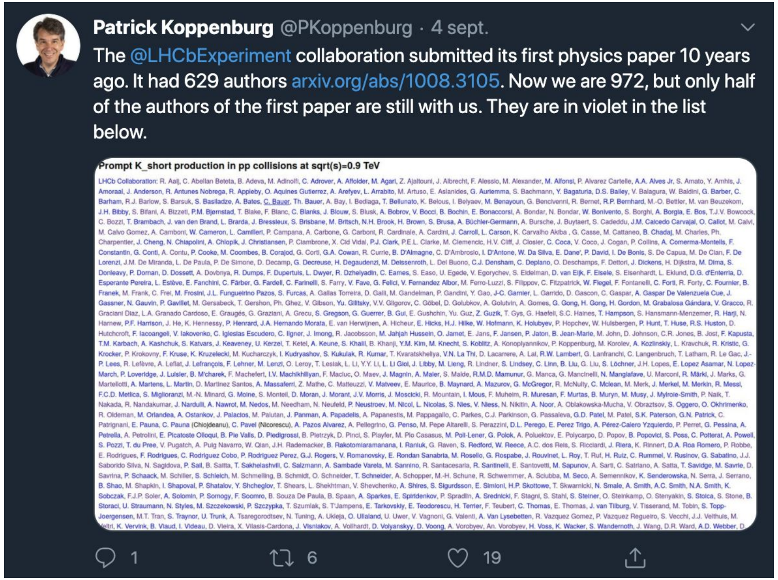


https://cds.cern.ch/record/910381

Data and knowledge

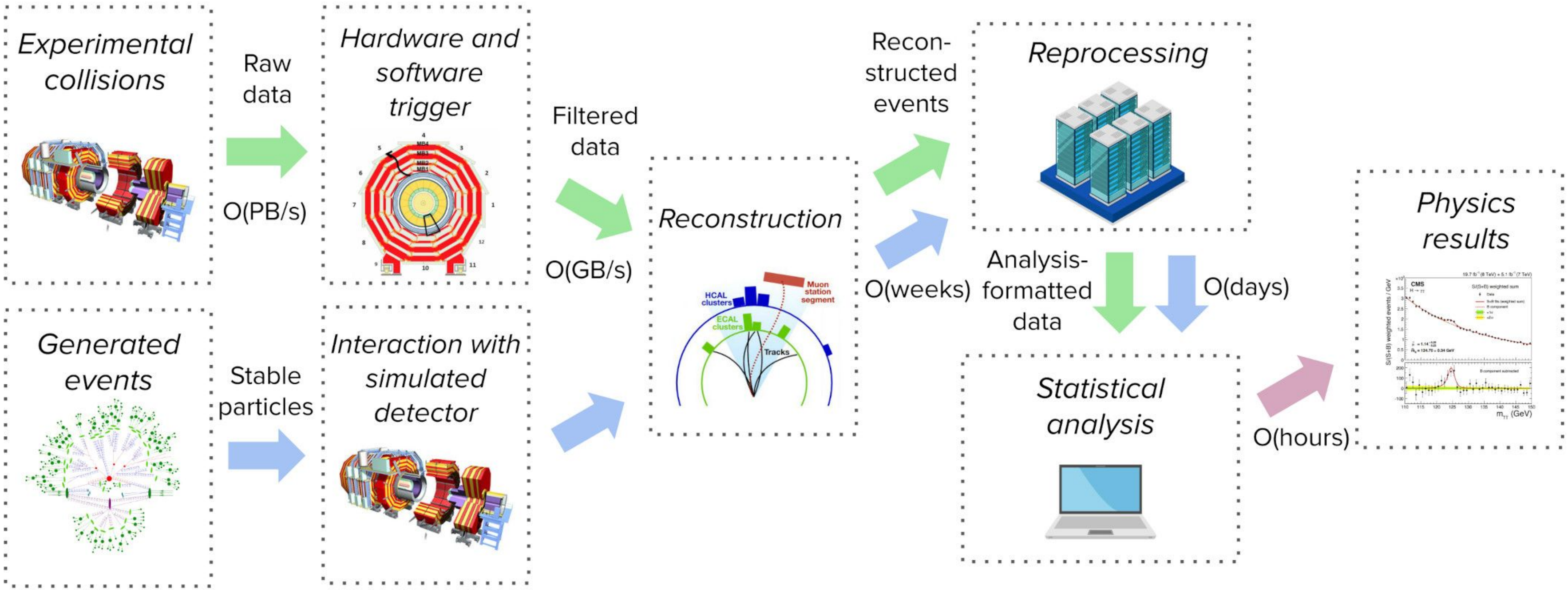


Achim Geiser https://indico.cern.ch/event/588219



https://twitter.com/PKoppenburg/status/1301813341460066304

HEP data analyses



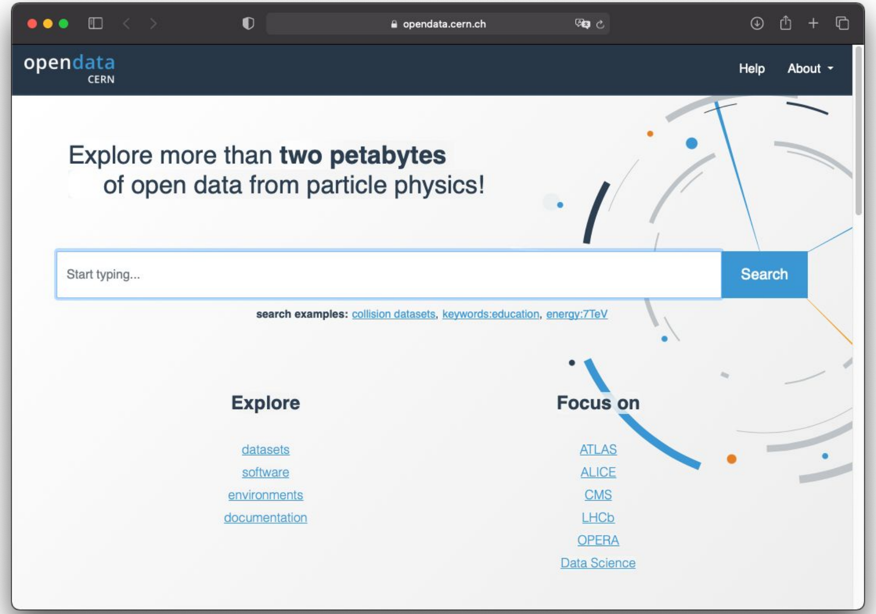
CERN Open Data portal

Launched in 2014

Disseminating over 2.4 PB of data

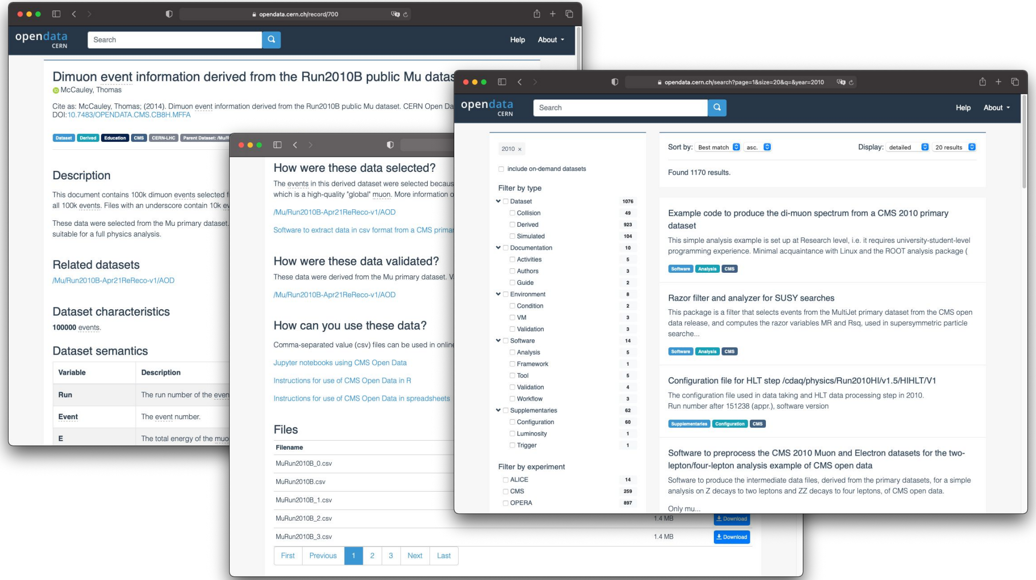
7.500 records

900.000 files



http://opendata.cern.ch/

Information organisation



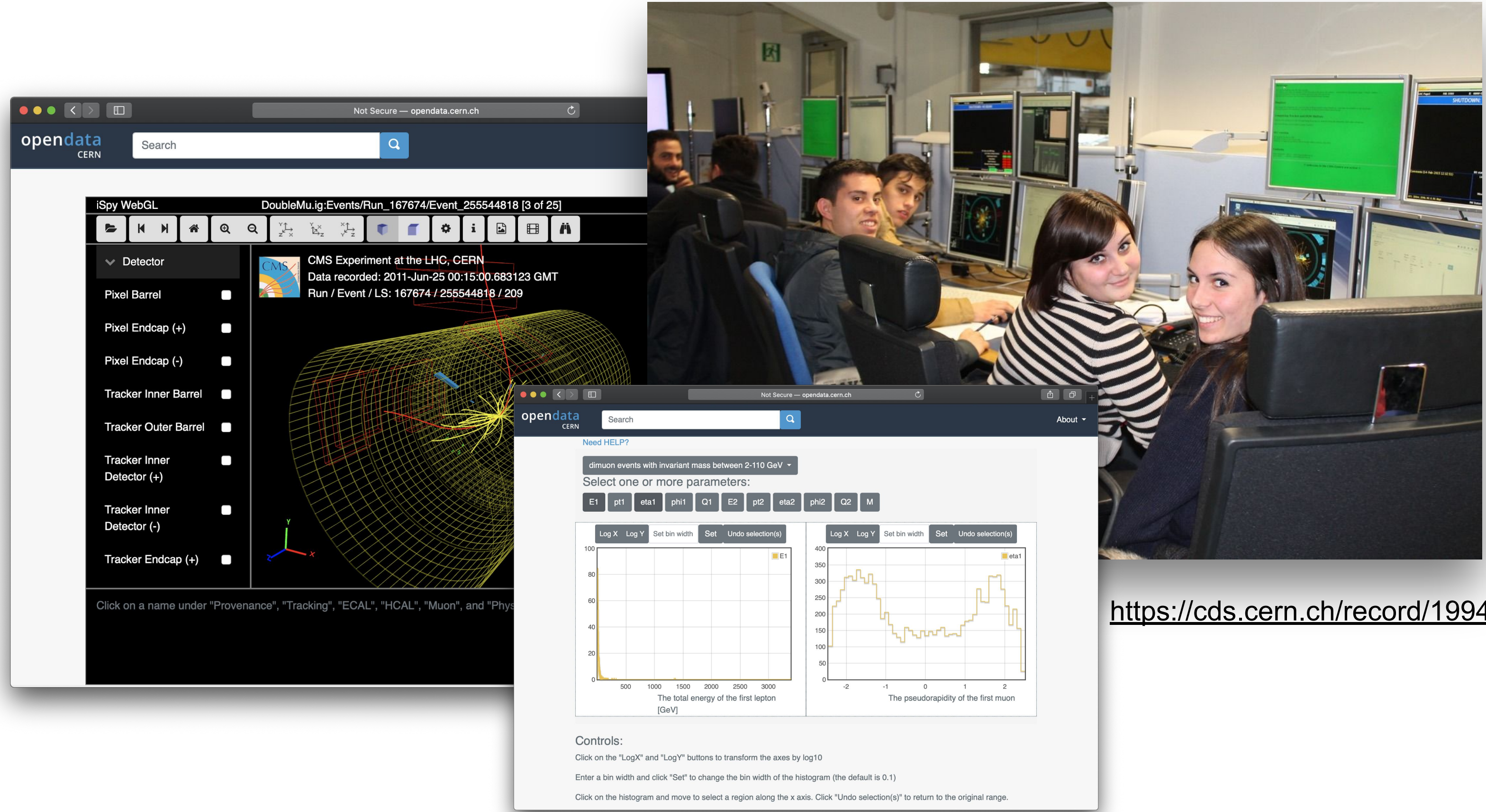
Faceted search

Large file download O(GB)

JSON Schema

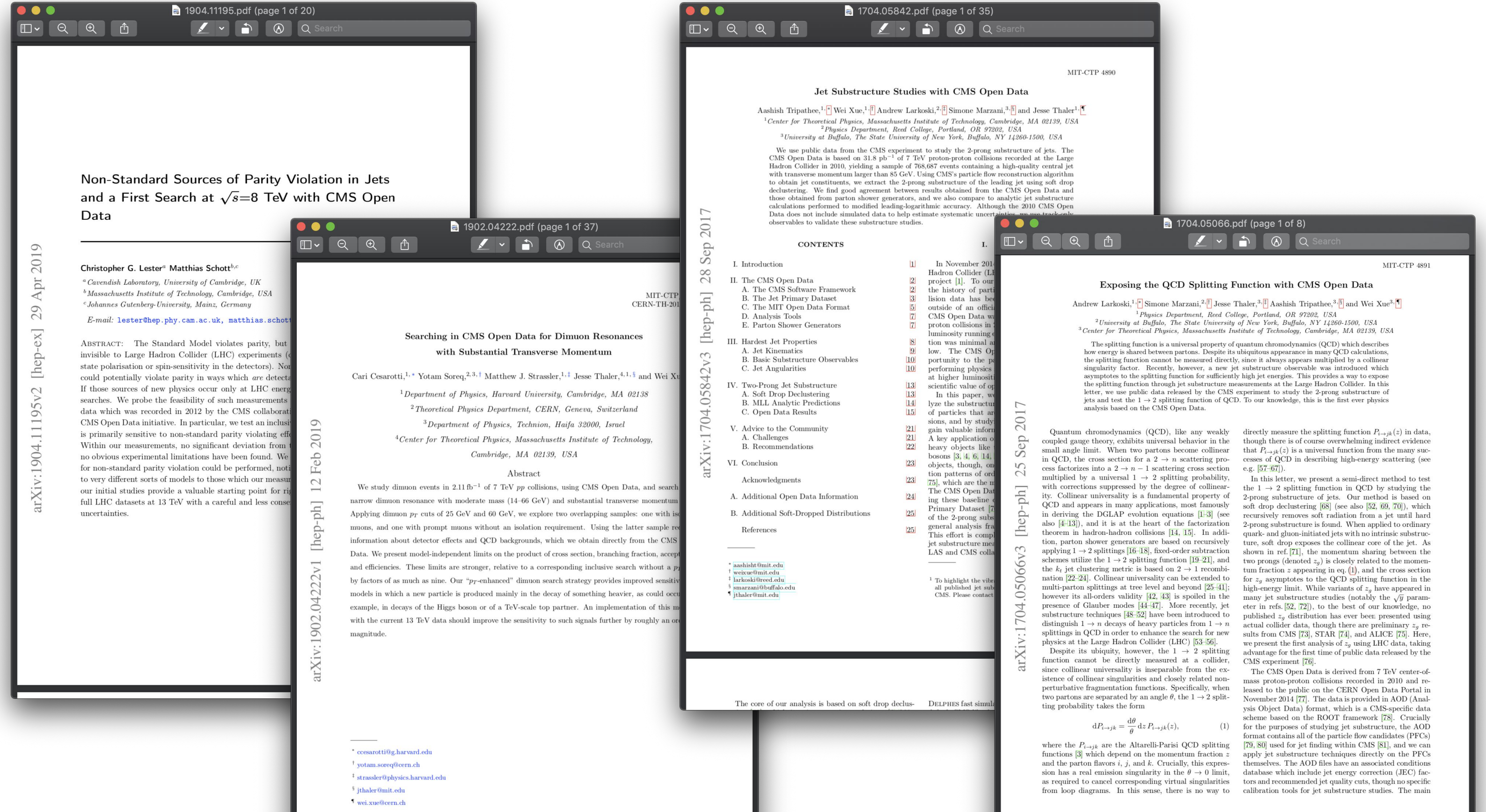
Data provenance

Education use cases



https://cds.cern.ch/record/1994217

Research use cases



Analysis examples

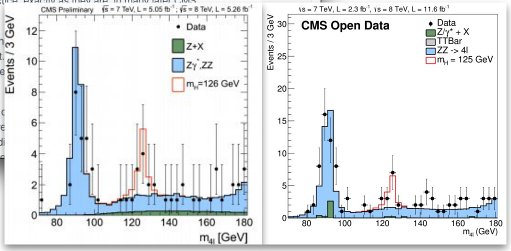
code

+

data

+

environment



http://opendata.cern.ch/record/5500

The four questions

where is **data**?

hard drive, distributed storage

where is the **code**?

GitLab, local copy, email

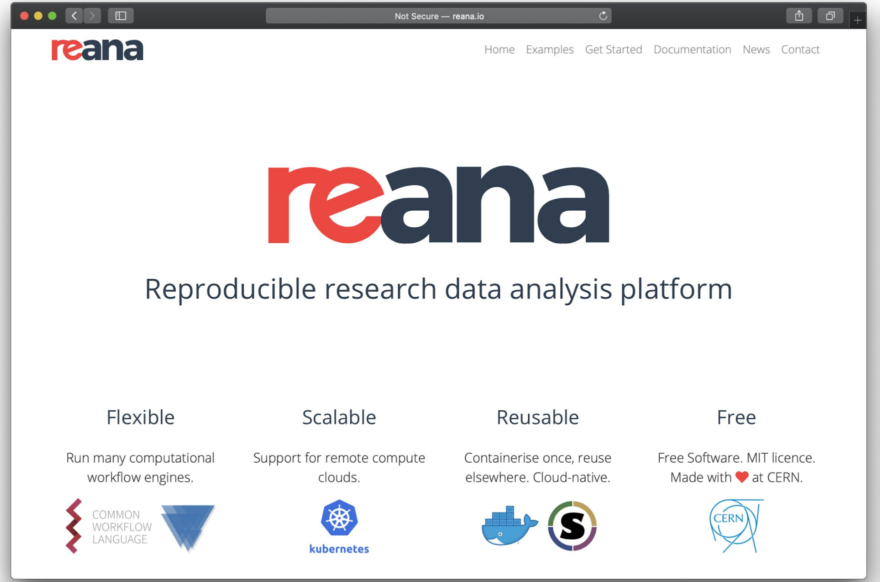
what **environment** do you use?

my own laptop, remote server

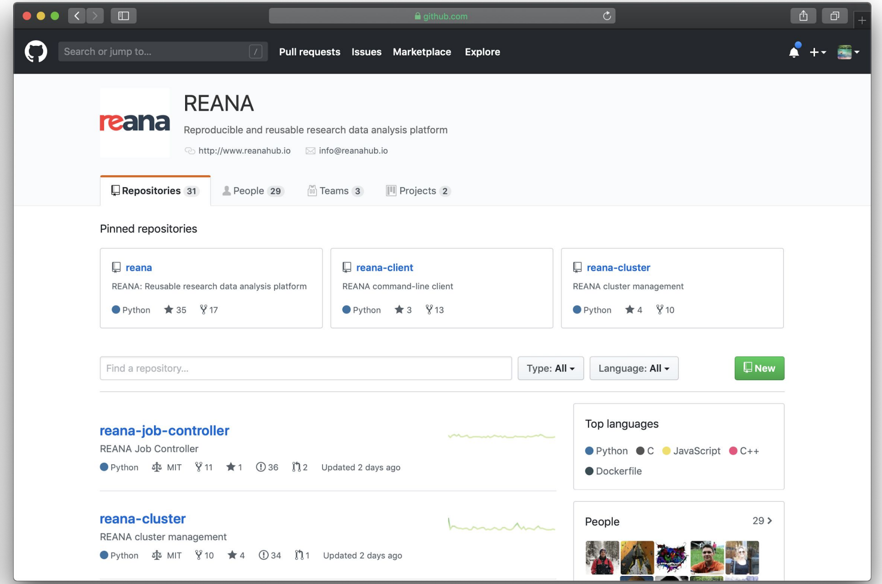
what **workflow** do you use**?**

Interactive commands, bash script, README file

Reproducible analyses



http://reana.io/



https://github.com/reanahub

Example

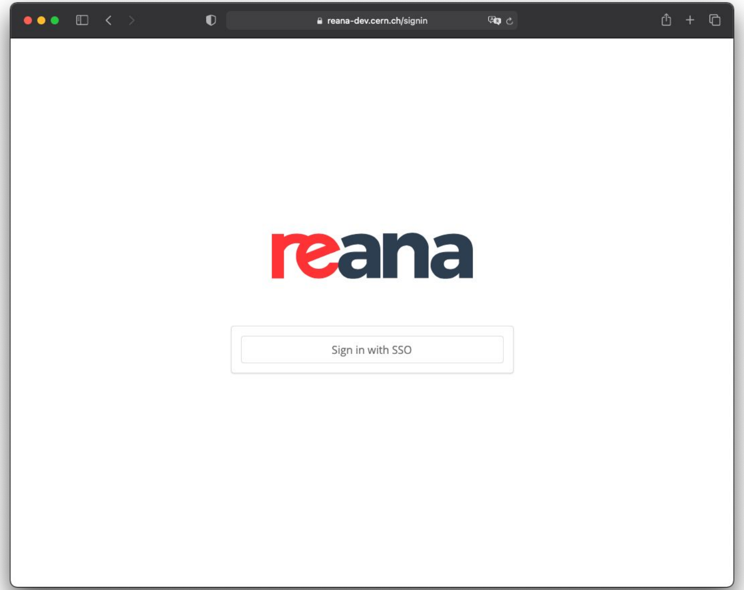
```
version: 0.6.0
inputs:
  files:
    - code/gendata.C
    - code/fitdata.C
  parameters:
    events: 20000
    data: results/data.root
    plot: results/plot.png
workflow:
  type: serial
  specification:
    steps:
      - name: gendata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
        - mkdir -p results && root -b -q
'code/gendata.C(${events},"${data}")'
      - name: fitdata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
        - root -b -q 'code/fitdata.C("${data}","${plot}")'
outputs:
  files:
    - results/plot.png
```

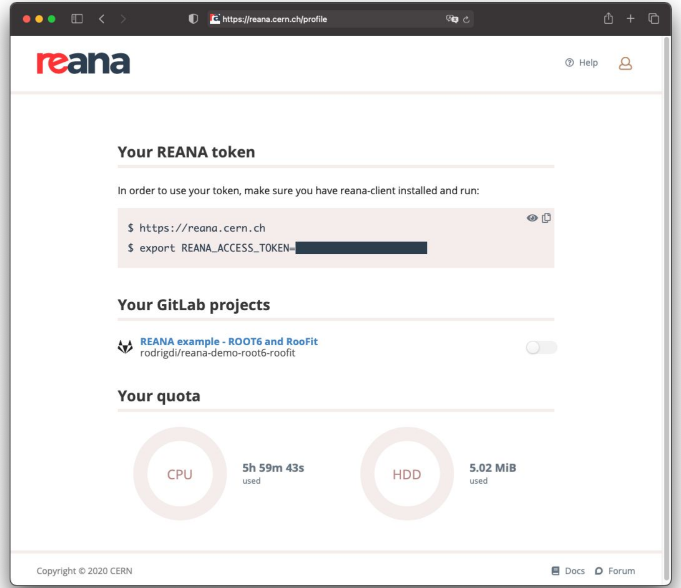
Inputs including: **code**, **data**, **parameters**

Workflow including: **steps/commands**, **container images**

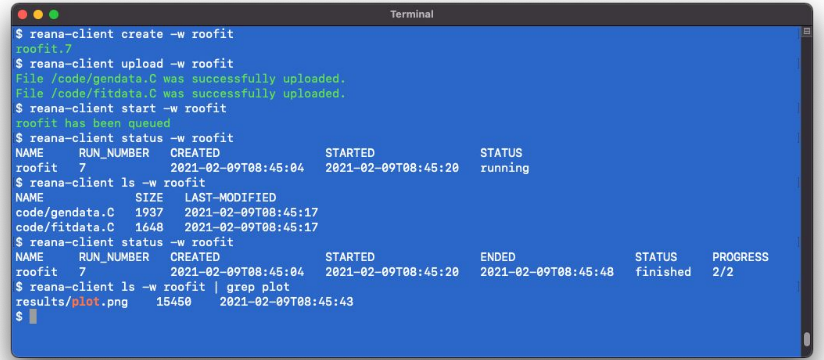
Outputs including: **plots**, **data etc...**

Example

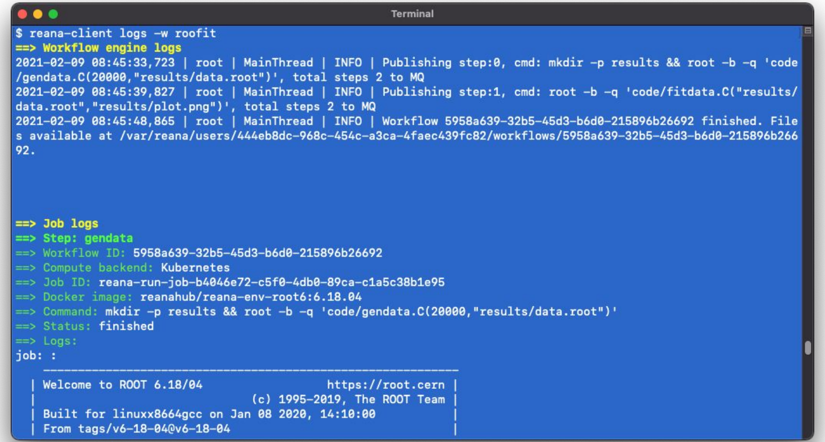

reana-dev.cern.ch/signin
reana
Sign in with SSO


https://reana.cern.ch/profile
reana
Help
Your REANA token
In order to use your token, make sure you have reana-client installed and run:
$ https://reana.cern.ch
$ export REANA_ACCESS_TOKEN=[redacted]
Your GitLab projects
REANA example - ROOT6 and RooFit
rodrigdi/reana-demo-root6-roofit
Your quota
CPU
5h 59m 43s
used
HDD
5.02 MiB
used
Copyright © 2020 CERN
Docs
Forum

Example

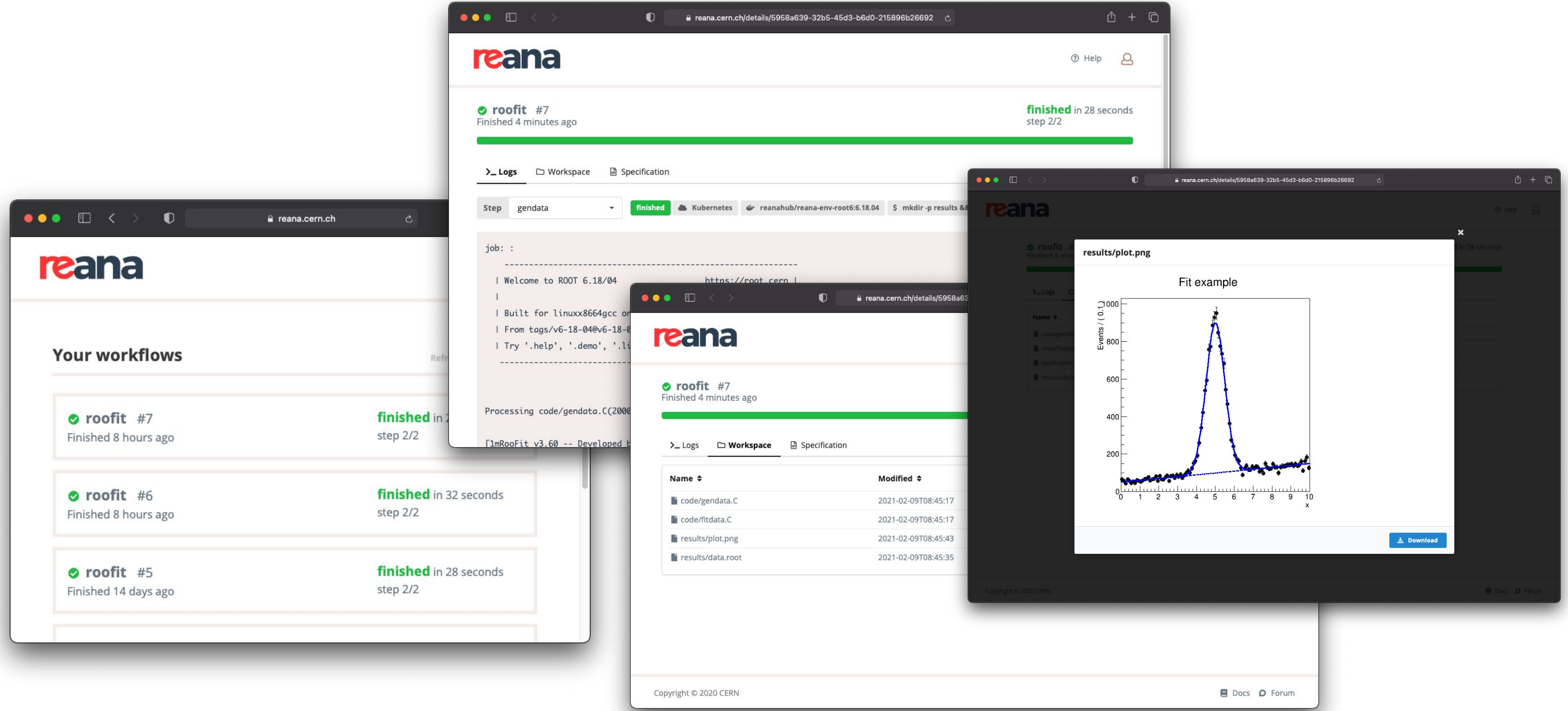


Terminal
$ reana-client create -w roofit
roofit.7
$ reana-client upload -w roofit
File /code/gendata.C was successfully uploaded.
File /code/fitdata.C was successfully uploaded.
$ reana-client start -w roofit
roofit has been queued
$ reana-client status -w roofit
NAME RUN_NUMBER CREATED STARTED STATUS
roofit 7 2021-02-09T08:45:04 2021-02-09T08:45:20 running
$ reana-client ls -w roofit
NAME SIZE LAST-MODIFIED
code/gendata.C 1937 2021-02-09T08:45:17
code/fitdata.C 1648 2021-02-09T08:45:17
$ reana-client status -w roofit
NAME RUN_NUMBER CREATED STARTED ENDED STATUS PROGRESS
roofit 7 2021-02-09T08:45:04 2021-02-09T08:45:20 2021-02-09T08:45:48 finished 2/2
$ reana-client ls -w roofit | grep plot
results/plot.png 15450 2021-02-09T08:45:43
$

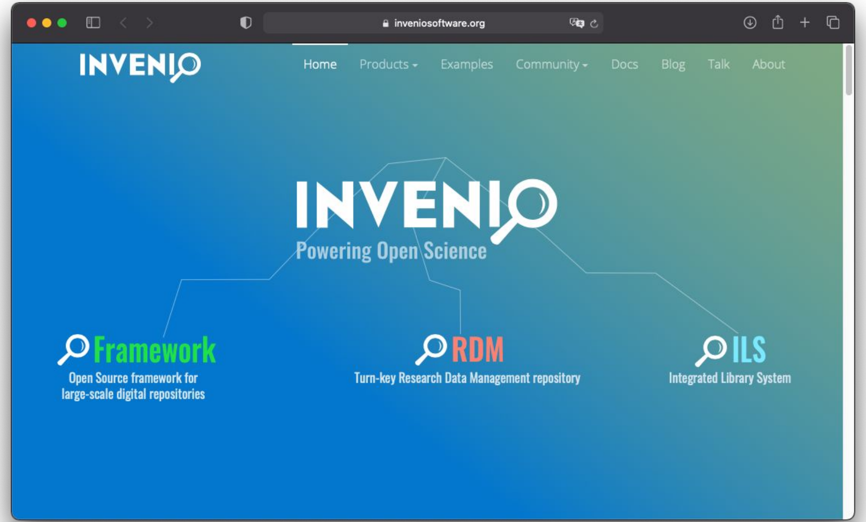


Terminal
$ reana-client logs -w roofit
==> Workflow engine logs
2021-02-09 08:45:33,723 | root | MainThread | INFO | Publishing step:0, cmd: mkdir -p results && root -b -q 'code/gendata.C(20000,"results/data.root")', total steps 2 to MQ
2021-02-09 08:45:39,827 | root | MainThread | INFO | Publishing step:1, cmd: root -b -q 'code/fitdata.C("results/data.root","results/plot.png")', total steps 2 to MQ
2021-02-09 08:45:48,865 | root | MainThread | INFO | Workflow 5958a639-32b5-45d3-b6d0-215896b26692 finished. Files available at /var/reana/users/444eb8dc-968c-454c-a3ca-4faec439fc82/workflows/5958a639-32b5-45d3-b6d0-215896b26692.
==> Job logs
==> Step: gendata
==> Workflow ID: 5958a639-32b5-45d3-b6d0-215896b26692
==> Compute backend: Kubernetes
==> Job ID: reana-run-job-b4046e72-c5f0-4db0-89ca-c1a5c38b1e95
==> Docker image: reanahub/reana-env-root6:6.18.04
==> Command: mkdir -p results && root -b -q 'code/gendata.C(20000,"results/data.root")'
==> Status: finished
==> Logs:
job: :
| Welcome to ROOT 6.18/04 https://root.cern |
| (c) 1995-2019, The ROOT Team |
| Built for linuxx8664gcc on Jan 08 2020, 14:10:00 |
| From tags/v6-18-04@v6-18-04 |

Example


reana.cern.ch
reana
Your workflows
roofit #7
Finished 8 hours ago
finished
step 2/2
roofit #6
Finished 8 hours ago
finished in 32 seconds
step 2/2
roofit #5
Finished 14 days ago
finished in 28 seconds
step 2/2
reana.cern.ch/details/5958a639-32b5-45d3-b6d0-215896b26692
Help
roofit #7
Finished 4 minutes ago
finished in 28 seconds
step 2/2
Logs
Workspace
Specification
Step
gendata
finished
Kubernetes
reanahub/reana-env-root6:6.18.04
$ mkdir -p results &&
job: :
Welcome to ROOT 6.18/04
https://root.cern
Built for linuxx8664gcc
From tags/v6-18-04@v6-18-
Try '.help', '.demo', '.li
Processing code/gendata.C(2000
[1mRooFit v3.60 -- Developed
Name
Modified
code/gendata.C
2021-02-09T08:45:17
code/fitdata.C
2021-02-09T08:45:17
results/plot.png
2021-02-09T08:45:43
results/data.root
2021-02-09T08:45:35
Copyright © 2020 CERN
Docs
Forum
results/plot.png
Fit example
Events / (0.1)
x
Download

Technology: repository



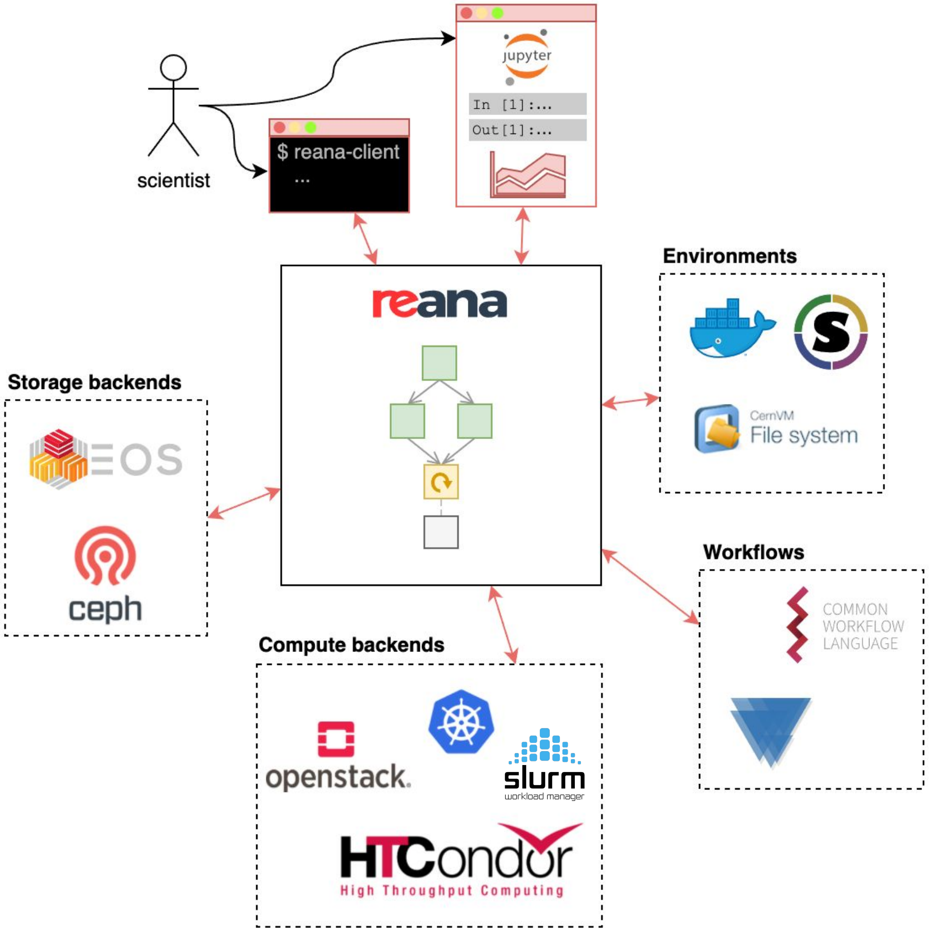
https://inveniosoftware.org

The technology behind:
https://www.hepdata.net/
https://inspirehep.net/
https://zenodo.org

- Python
- Flask
- React
- JSONSchema
- Redis
- Elasticsearch
- PostgreSQL

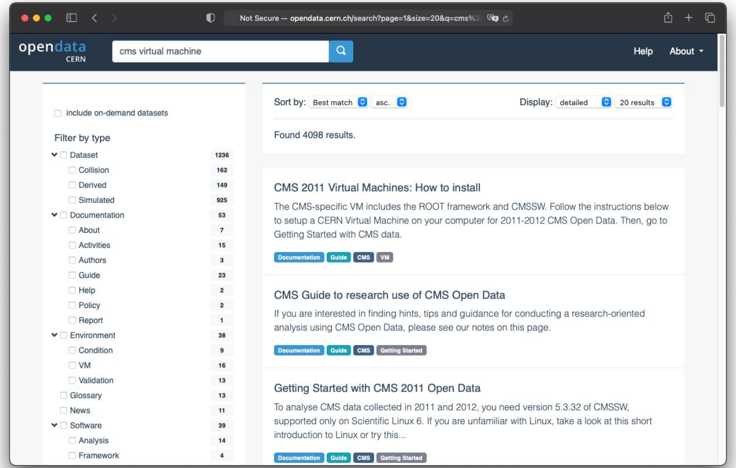
Technology: REANA

- Cloud-native application
- Extensible
  - Storage backends
  - Compute backends
  - Container technologies
  - Workflow engines



Try them out!

Use opendata.cern.ch



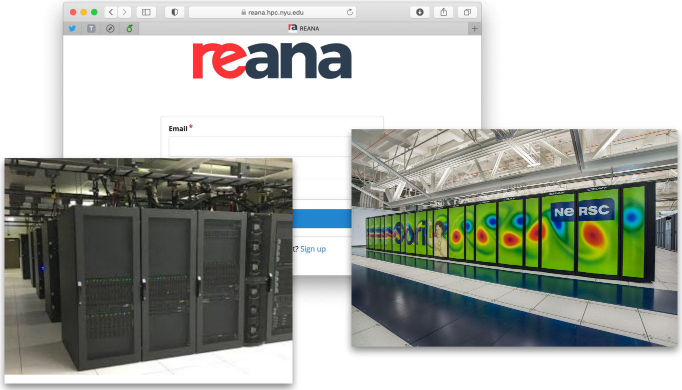
Install REANA on premises/locally

```
$ helm repo add reanahub \
  https://reanahub.github.io/reana

$ helm repo update

$ helm install reana reanahub/reana
```

Helm repository at http://reanahub.github.io/reana, more documentation at http://docs.reana.io/development/deploying-at-scale/



What's next

Roadmap

Near-term

What we plan to work on next

Live logs

Introduce live job log streaming for CLI and Web UI.

LHC community

Introduce abstract dataset concept to handle a set of related files.
Use various remote storage backends for workflow workspace.

Future

What is coming later

Abstraction of data storage

Use various remote storage backends for workflow workspace.

User groups and authorisations

Introduce OpenID Connect to support more authentication mechanisms.
Introduce user groups and role-based authorisation control models.

Get in touch

CERN Open Data

- https://opendata.cern.ch
- https://github.com/cernopendata
- https://forum.opendata.ch/
- https://gitter.im/cernopendata/opendata.cern.ch
- https://twitter.com/cernopendata

REANA

- https://www.reana.io
- https://github.com/reanahub/reana
- https://forum.reana.io/
- https://gitter.im/reanahub/reana
- https://twitter.com/reanahub
- https://docs.reana.io