Virtual Research Environment goes live:
Big data shared across scientific analyses

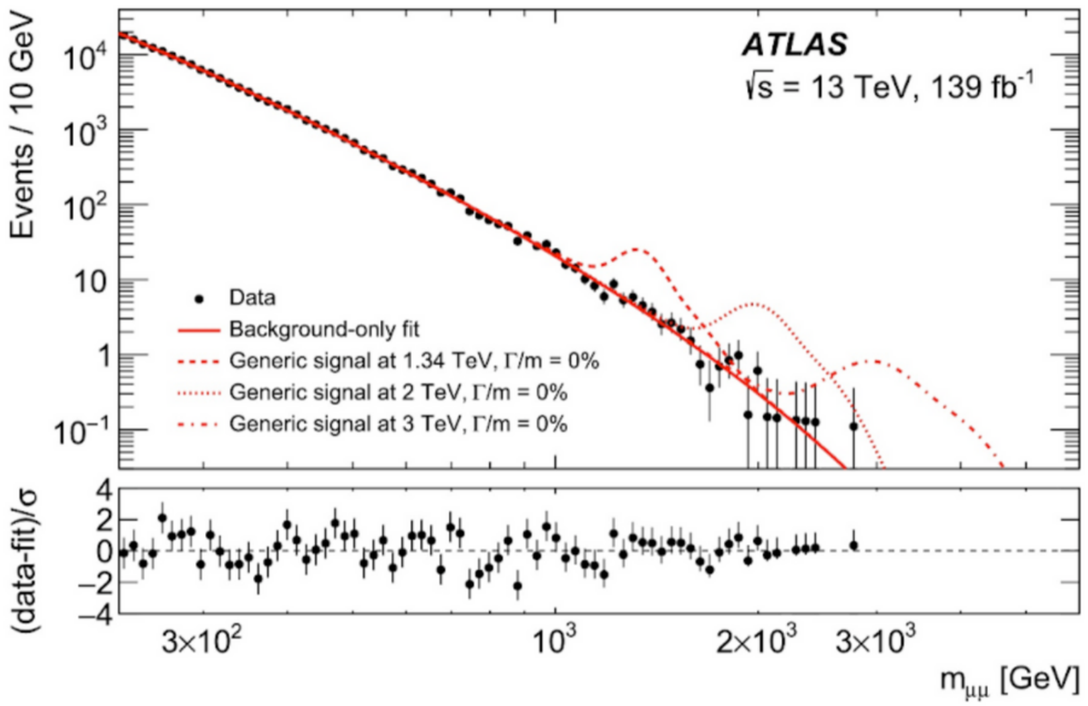
Agisilaos Kounelis

Supervisors: Elena Gazzarrini, Alba Vendrell Moya, Xavier Espinal Curull
Contributor: Tibor Šimko

September 5, 2022
IT-GOV-ENG section meeting

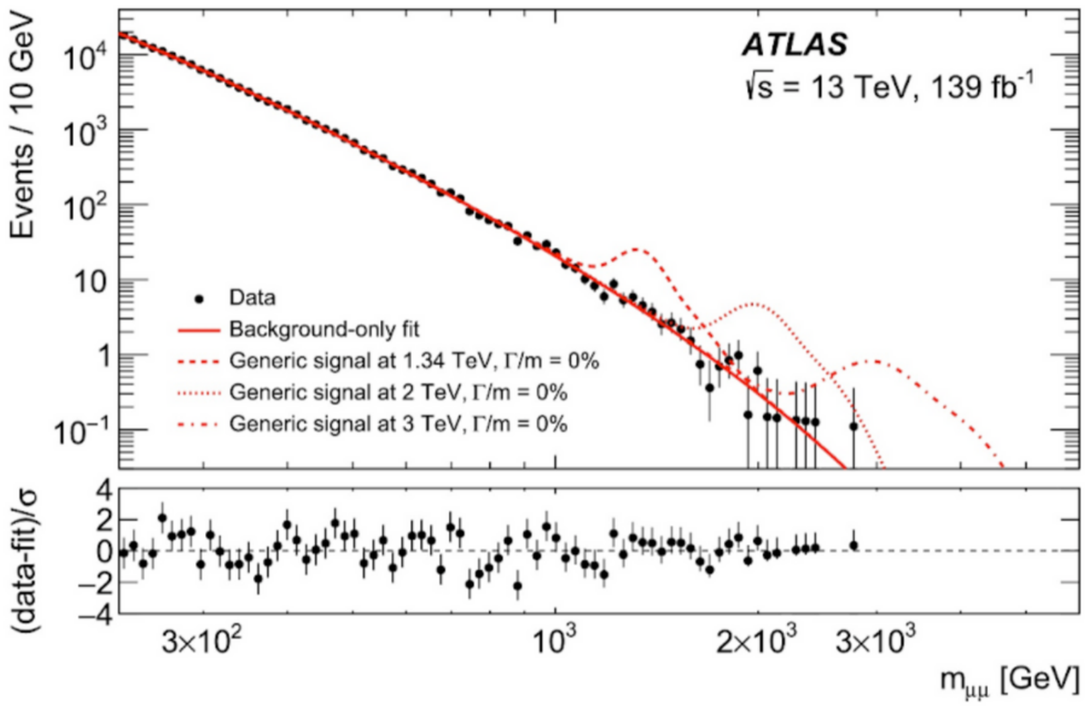
Introduction

- OPEN SOURCE SCIENCE
- COLLABORATION
- FAIR DATA PRINCIPLES
- ANALYSIS PRESERVATION



Introduction

- OPEN SOURCE SCIENCE
- COLLABORATION
- FAIR DATA PRINCIPLES
- ANALYSIS PRESERVATION



Need for appropriate analysis tools

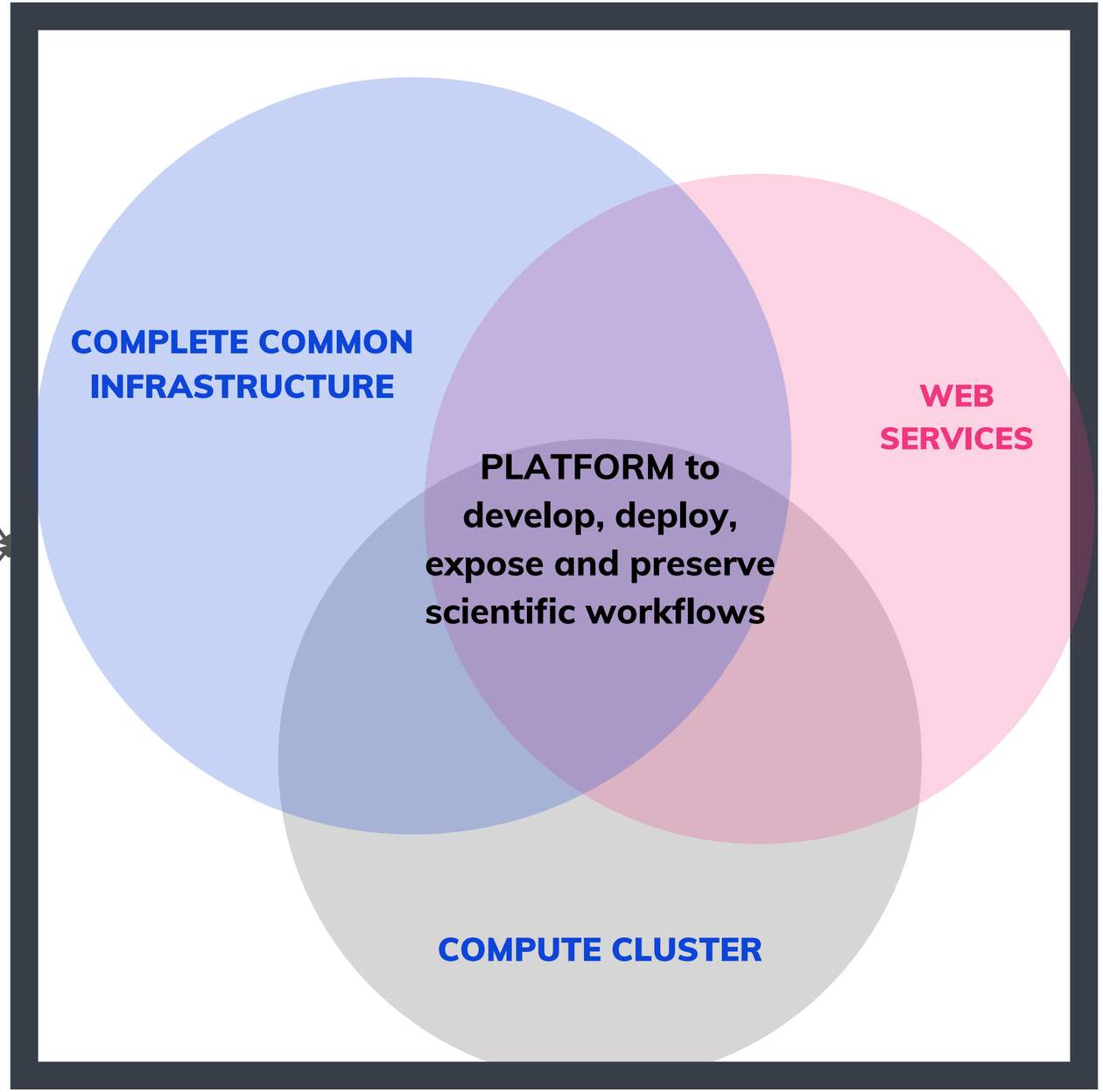
EOSC-FUTURE

DARK MATTER

LHC
Neutrino Observations
Indirect Detection Methods

EXTREME UNIVERSE

Compact objects
Active Galactic Nuclei
Multimessenger Astronomy



different data needs

+

different methodologies

VRE

Generation and simulation of events

Experimental DATA from ESCAPE experiments

Data processing

Including reconstruction and calibration path

Analysis

Including background subtraction

+

Interpretation of results

COMBINATION of results from different experiments

COMPARISON of results with other searches

KM3NeT, ATLAS EXPERIMENT, DARKSIDE, ASTRON, EGO, cta

FTS, kubernetes, RUCIO

ESCAPE's Data Lake running on a K8s cluster

jupyter

WebUI: DataLake-as-a-Service

+

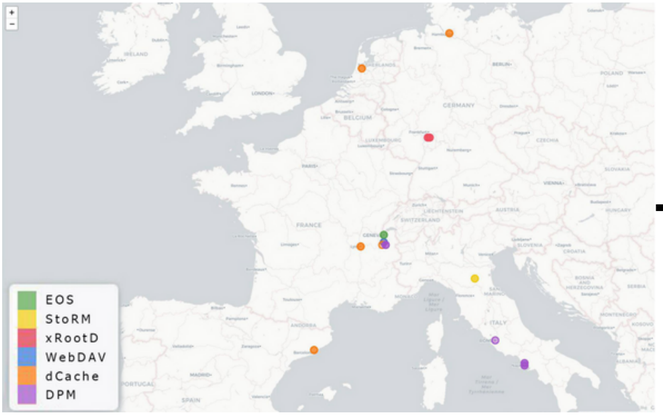
reana

Platform to reproduce workflows running on various computing backends

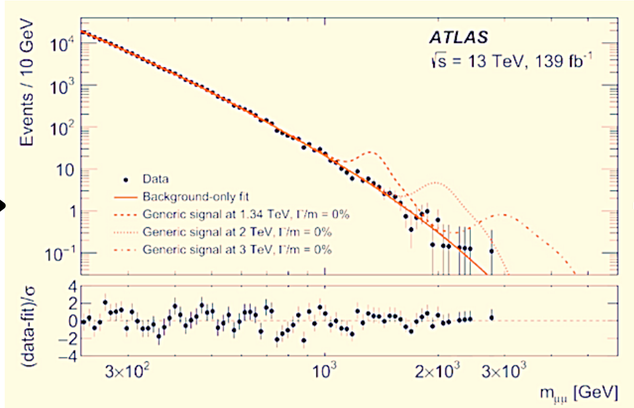
CERN openlab, ESCAPE, CERN, EOSC Future

Problem

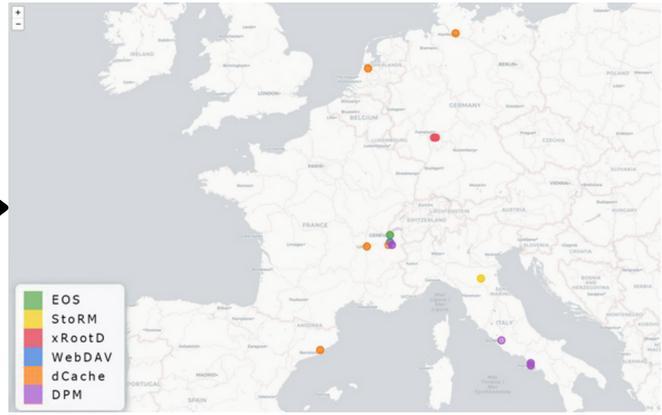
Manage plenty of Data spread across facilities at multiple institutions



Analyze and Extract useful conclusions from big data



Save Results back into data facilities



Rucio

SCIENTIFIC DATA MANAGEMENT

RUCIO

What?

- CERN service that manages large volumes of data spread across facilities

Why?

- Centralises the control of all data
- Highly scalable and modular
- Orchestrates data sharing

Who?

- High Energy Physics as the main target
- Becoming popular within diverse scientific communities (e.g. astronomy, biology).

REANA

What?

- CERN's re-analysis platform allowing researchers to run remote analyses.

Why?

- powerful - takes advantage of declarative (.yaml files) instead of imperative
- portability - thanks to recent advances of cloud container technology

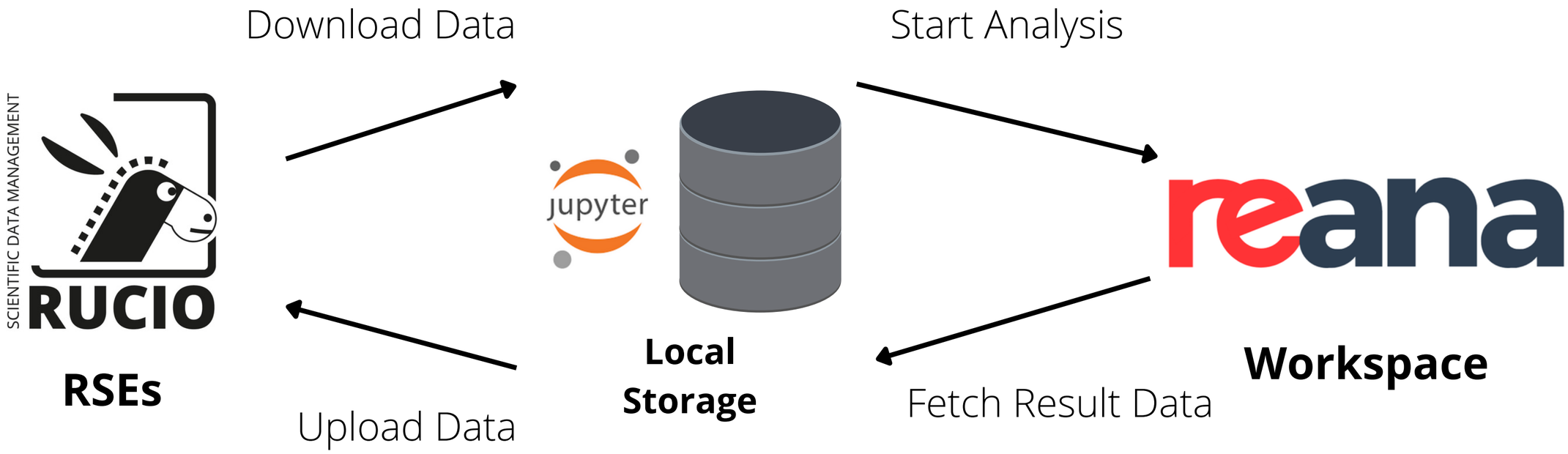
How?

- Input data (files, parameters)
- Analysis code (user code, software code)
- Computing environment (operating system, database calls)
- Computational recipes (shell commands, notebooks and workflows)

Who?

- Was born for particle physics analyses, but applicable to any scientific discipline.

Existing Solution



Existing Solution - Step 1: Download data before starting analysis

SCIENTIFIC DATA MANAGEMENT

RUCIO

Terminal 1

```
jovyan@jupyter-akouneli:~/presentation$ rucio get agis_test:test.txt
  warnings.warn(
2022-08-31 15:37:45,331 INFO    Trying to download with davs and timeout of 60s from EULAKE-1: agis_test:test.txt
2022-08-31 15:37:45,332 INFO    Using PFN: davs://eoseulake.cern.ch:443/eos/eulake/tests/rucio_test/eulake_1/agis_test/75/5c/test.txt
/opt/conda/lib/python3.8/site-packages/urllib3/connectionpool.py:1013: InsecureRequestWarning: Unverified HTTPS request is being made to ho
st 'escape-rucio.cern.ch'. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usag
e.html#ssl-warnings
  warnings.warn(
2022-08-31 15:37:45,406 INFO    File agis_test:test.txt successfully downloaded. 11.000 B in 0.05 seconds = 0.0 MBps
----------------------------------
Download summary
----------------------------------------
DID agis_test:test.txt
Total files (DID):                          1
Total files (filtered):                     1
Downloaded files:                           1
Files already found locally:                0
Files that cannot be downloaded:            0
jovyan@jupyter-akouneli:~/presentation$
```

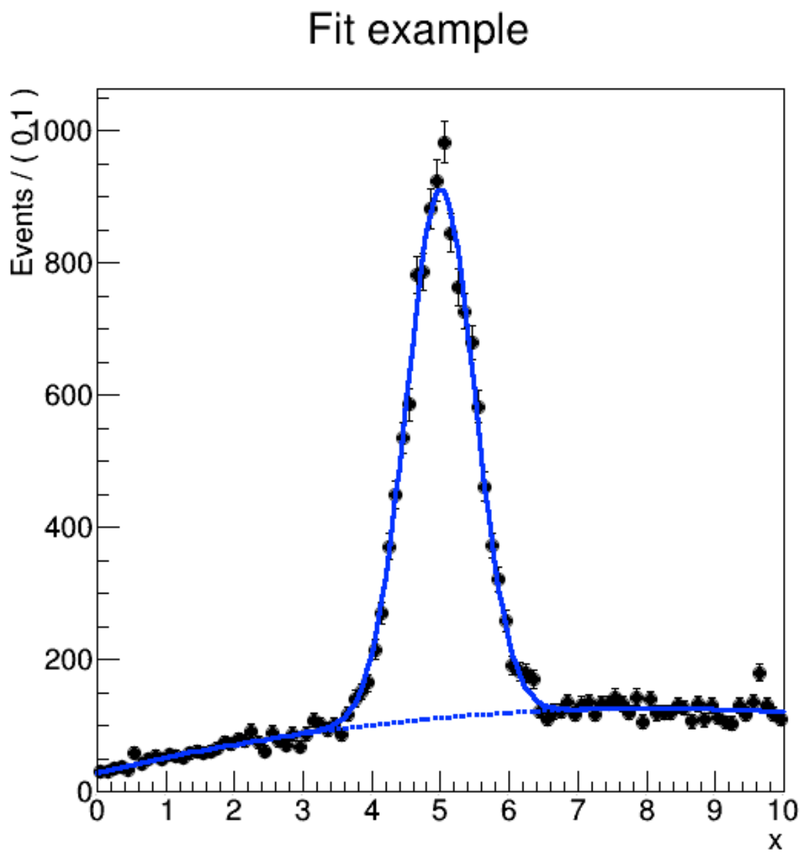
Existing Solution - Step 2: Run analysis with local files

reana

- Upload local data on REANA workspace
- Run analysis
- Store results on REANA workspace
- Download results locally

```
inputs:
  files:
    - code/gendata.C
    - code/fitdata.C
```

```
outputs:
  files:
    - results/plot.png
```



CERN openlab ESCAPE CERN EOSC Future

Existing Solution - Step 3: Upload data after analysis is completed

SCIENTIFIC DATA MANAGEMENT

RUCIO

Terminal 1

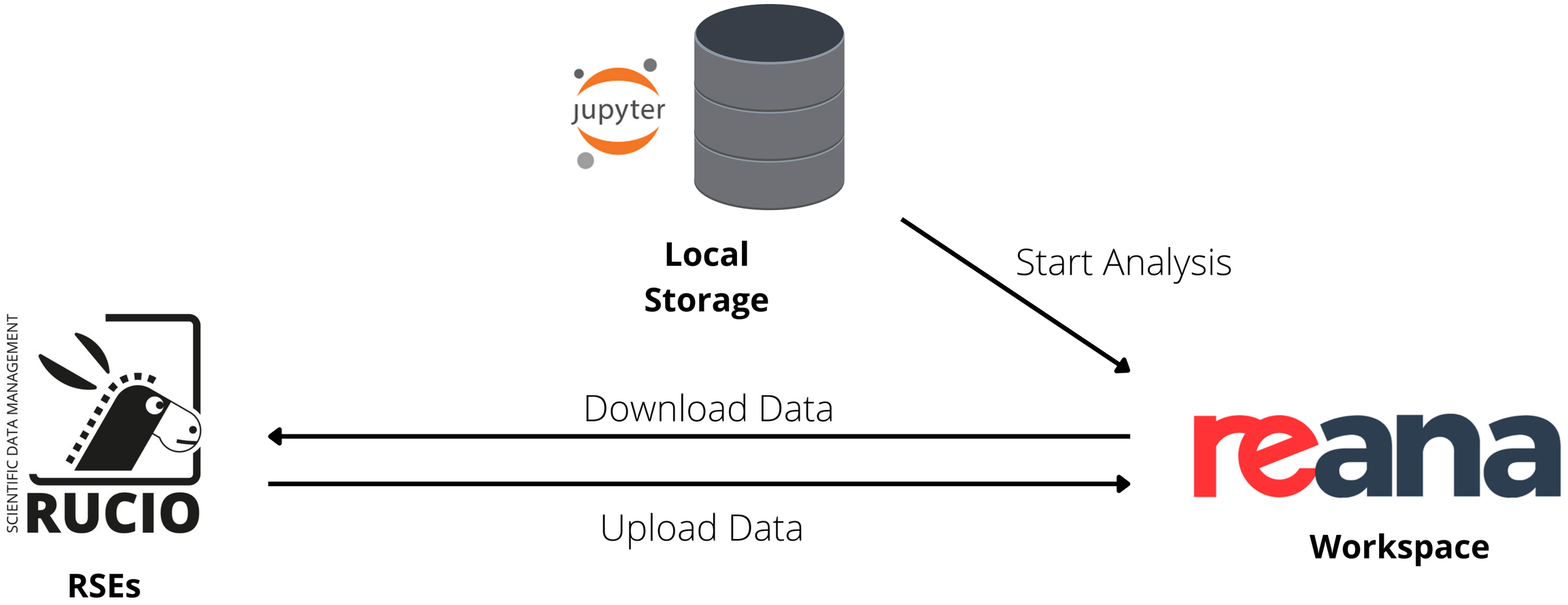
```
jovyan@jupyter-akouneli:~/presentation$ rucio upload --scope agis_test --rse EULAKE-1 test.txt
2022-08-31 15:35:13,019 INFO    Preparing upload for file test.txt
2022-08-31 15:35:13,368 INFO    Successfully added replica in Rucio catalogue at EULAKE-1
2022-08-31 15:35:13,594 INFO    Successfully added replication rule at EULAKE-1
2022-08-31 15:35:13,762 INFO    Trying upload with davs to EULAKE-1
2022-08-31 15:35:14,353 INFO    Successful upload of temporary file. davs://eoseulake.cern.ch:443/eos/eulake/tests/rucio_test/eulake_1/agis
_test/75/5c/test.txt.rucio.upload
2022-08-31 15:35:14,413 INFO    Successfully uploaded file test.txt
/opt/conda/lib/python3.8/site-packages/urllib3/connectionpool.py:1013: InsecureRequestWarning: Unverified HTTPS request is being made to ho
st 'escape-rucio.cern.ch'. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/1.26.x/advanced-usag
e.html#ssl-warnings
  warnings.warn(
jovyan@jupyter-akouneli:~/presentation$
```

Existing Solution - Problems

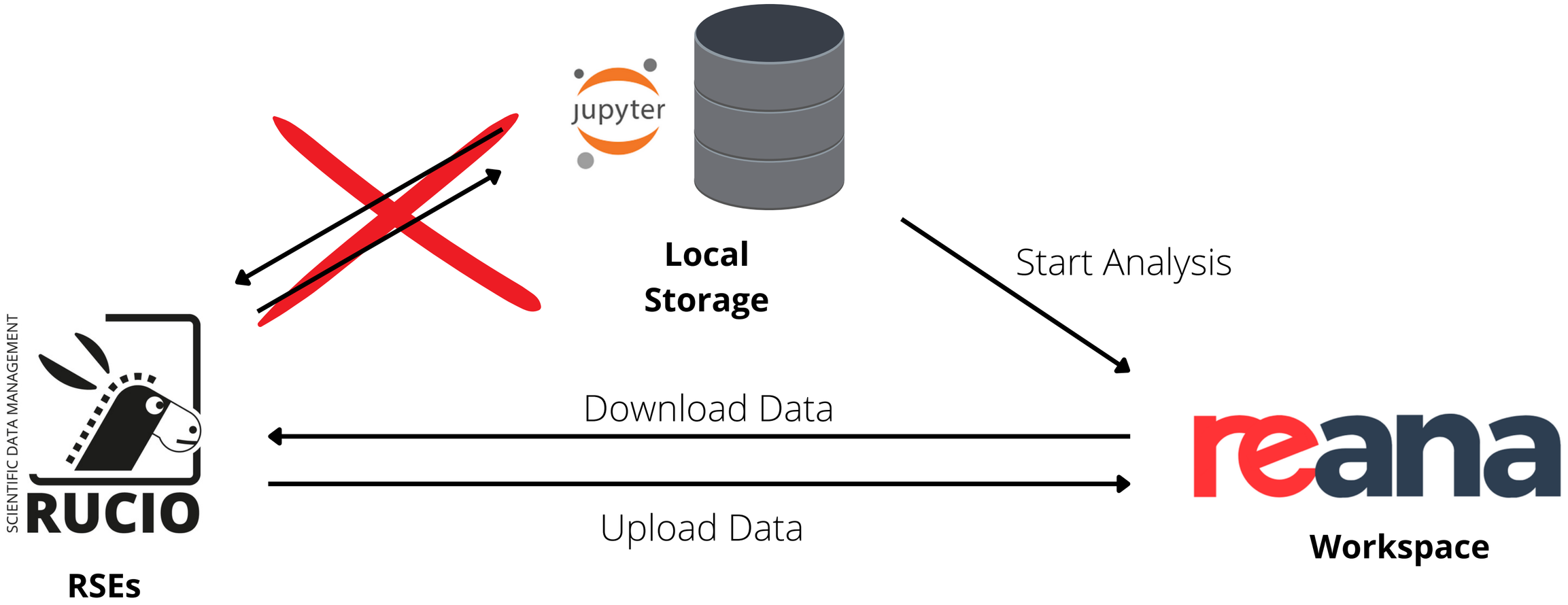
But, what if:

- Not enough available storage for the data
- Not good enough internet connection
- Someone else wants to execute the same workflow

Implemented Solution



Implemented Solution



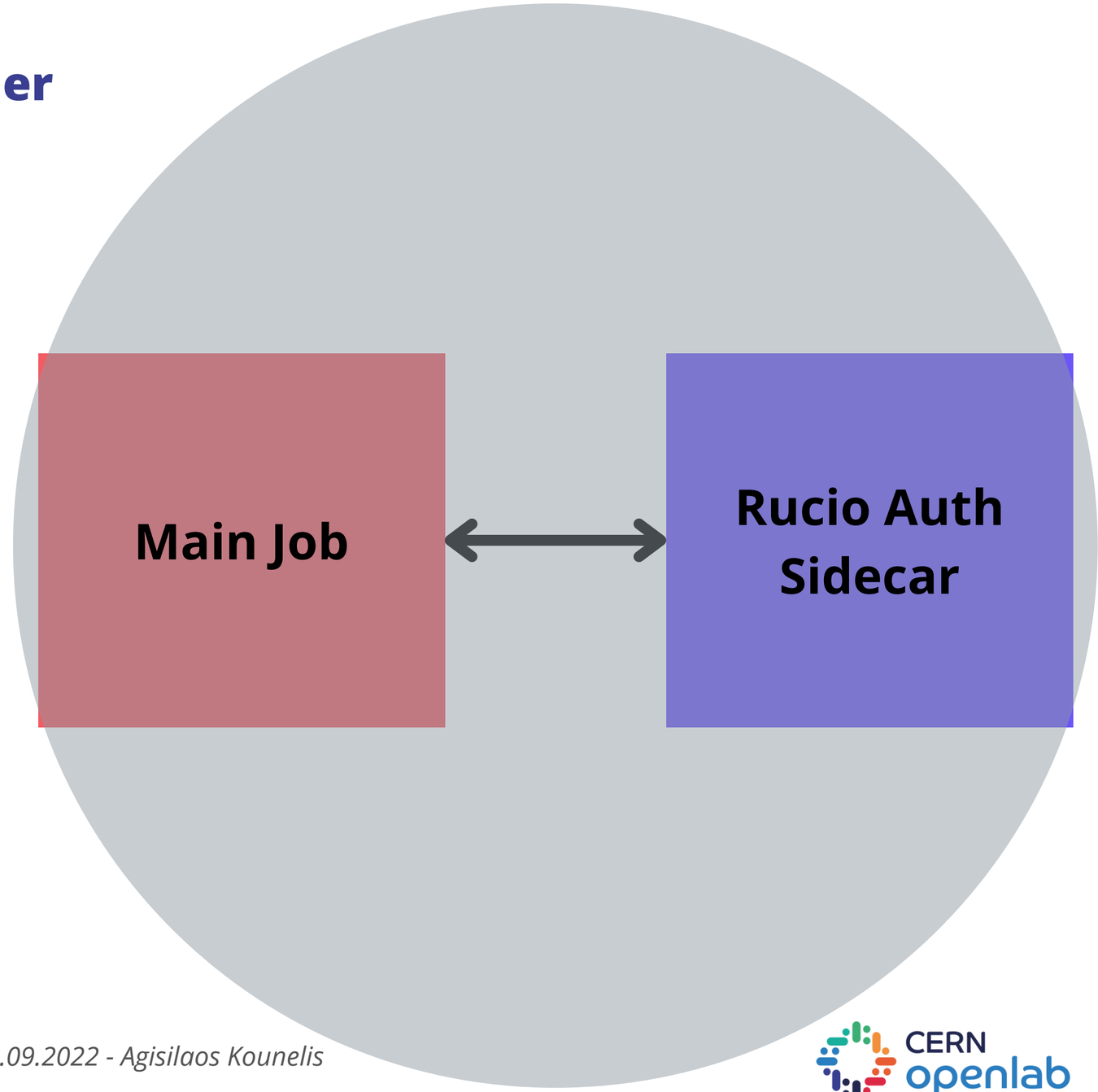
Contributions

Rucio Sidecar container	**REANA components**
Docker environments	**Documentation**

Contributions

Rucio Sidecar container	**REANA components**
Docker environments	**Documentation**

SIDECAR container



SIDECAR container

- Provide the appropriate configuration for Rucio
- Rucio authentication methods: OIDC, VOMS, x509, GSS, SSH key exchange, username/password, ...
- Selected **VOMS** authenication as it is already provided by REANA.

(1) Add secrets needed

$ reana-client secrets-add --file **userkey.pem** --file **usercert.pem** --env **VOMSPROXY_PASS**=XXXXXXX --env **VONAME**=escape --env **RUCIO_USERNAME**=akouneli

(2) Set **voms_proxy** & **rucio** to **true**

```
steps:
  - name: fetchdata
    voms_proxy: true
    rucio: true
    environment: 'reanahub/reana-env-rucioclient:latest'
    commands:
    - rucio get agis_test:fitdata.C agis_test:gendata.C
```

(3) Start of the workflow initializes the sidecar container

SIDECAR container

- Provide the appropriate configuration for Rucio
- Rucio authentication methods: OIDC, VOMS, x509, GSS, SSH key exchange, username/password, ...
- Selected **VOMS** authenication as it is already provided by REANA.

(1) Add secrets needed

$ reana-client secrets-add --file **userkey.pem** --file **usercert.pem** --env **VOMSPROXY_PASS**=XXXXXXX --env **VONAME**=escape --env **RUCIO_USERNAME**=akouneli

(2) Set **voms_proxy** & **rucio** to **true**

```
steps:
  - name: fetchdata
    voms_proxy: true
    rucio: true
    environment: 'reanahub/reana-env-rucioclient:latest'
    commands:
    - rucio get agis_test:fitdata.C agis_test:gendata.C
```

(3) Start of the workflow initializes the sidecar container.

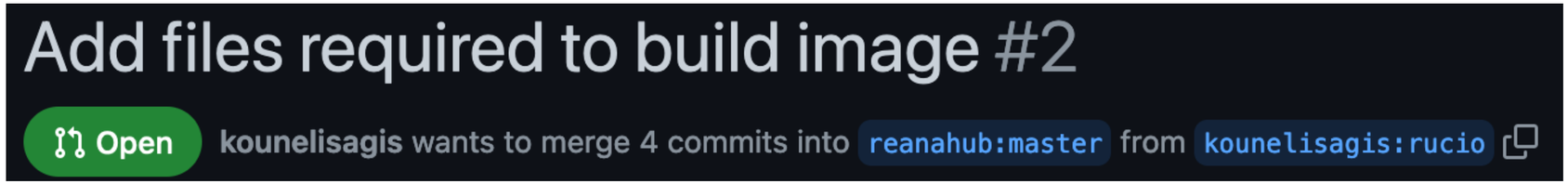
```
[client]
rucio_host = https://escape-rucio.cern.ch
auth_host = https://escape-rucio-auth.cern.ch
ca_cert = /rucio_cache/CERN-bundle.pem
auth_type = x509_proxy
account = akouneli
client_x509_proxy = /vomsproxy_cache/x509up_proxy
request_retries = 3
[policy]
permission = escape
schema = escape
lfn2pfn_algorithm_default = hash
support = https://github.com/rucio/rucio/issues/
support_rucio = https://github.com/rucio/rucio/issues/
```

SIDECAR container

Support for different Virtual Organizations (VOs)!

SIDECAR container

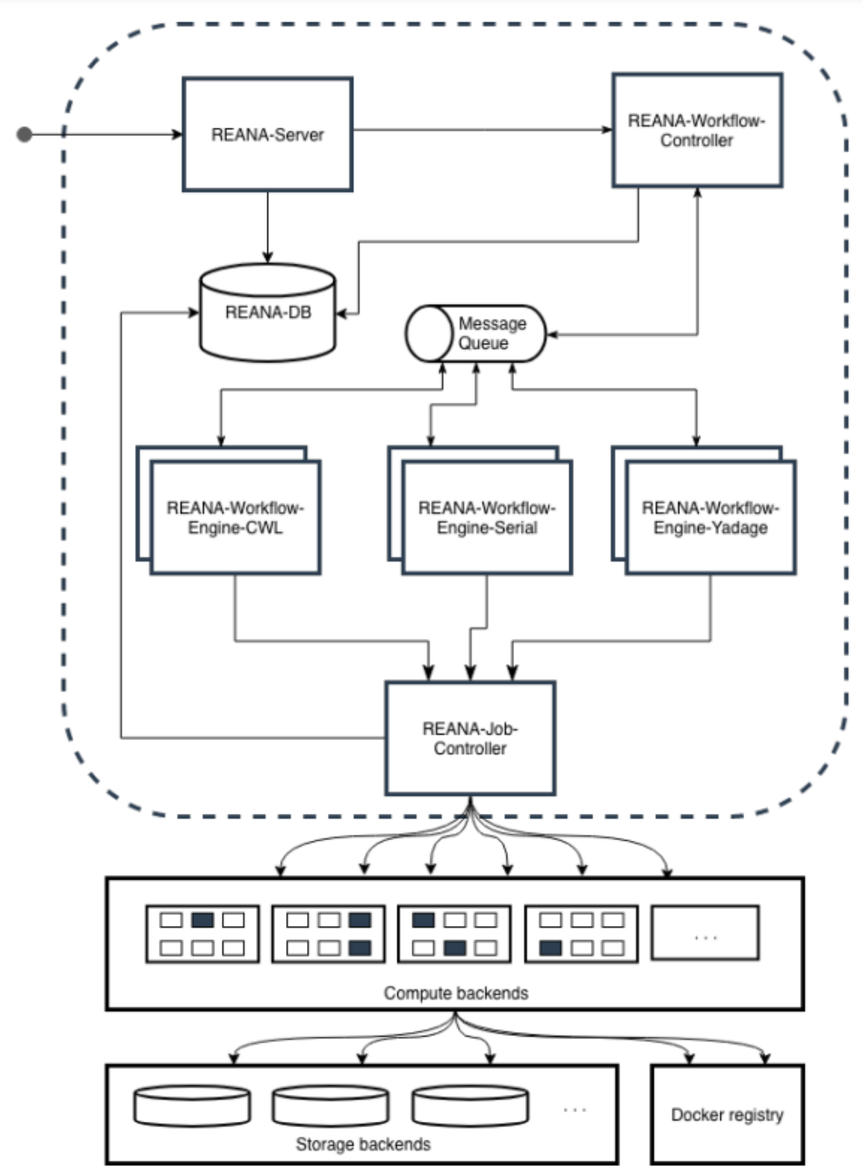
Support for different Virtual Organizations (VOs)!



Contributions

Rucio Sidecar container	REANA components
Docker environments	Documentation

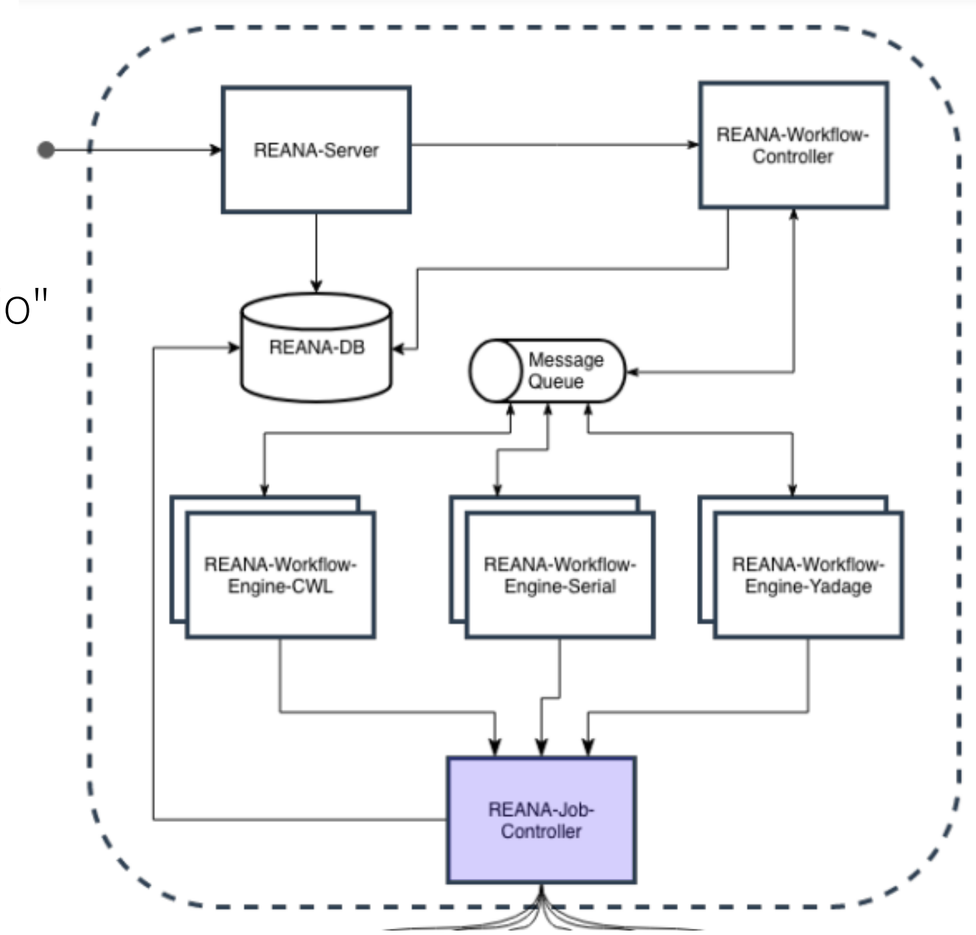
REANA components



REANA components

REANA-Job-Controller

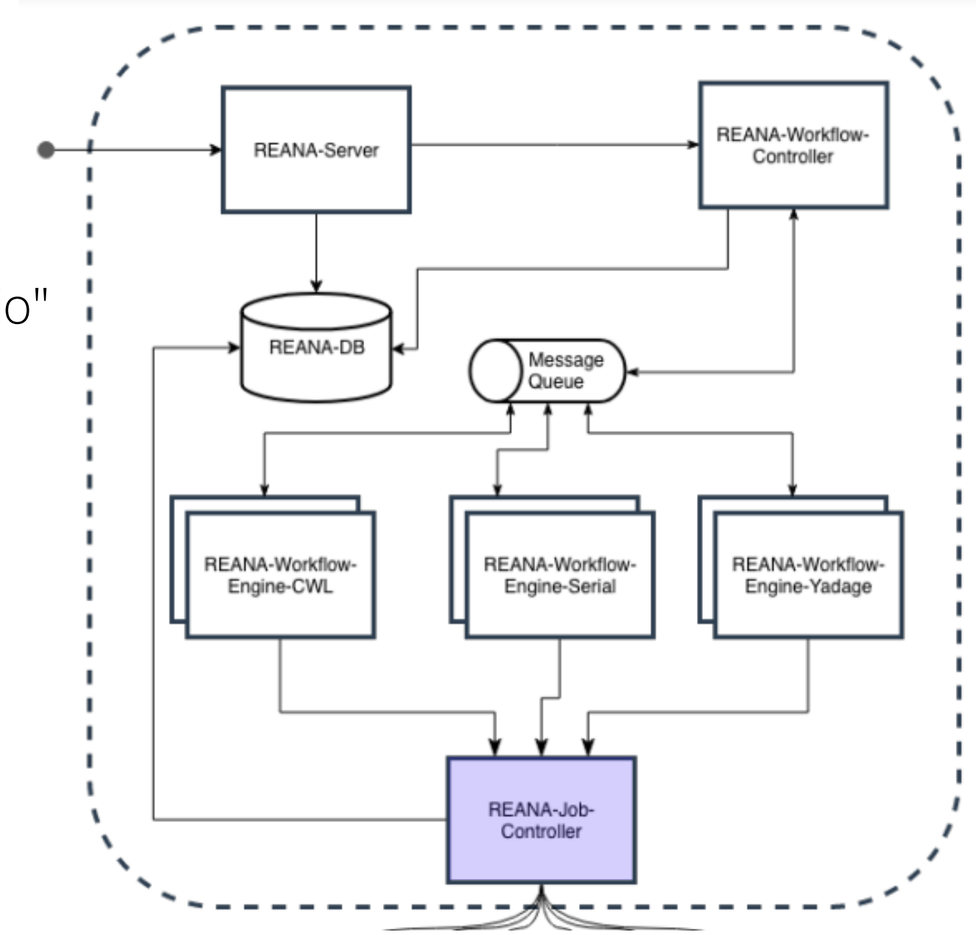
- Added Configuration:
  - RUCIO_CONTAINER_NAME = "reana-auth-rucio"
  - RUCIO_CACHE_LOCATION = "/rucio_cache/"
  - RUCIO_CFG_CACHE_FILENAME = "rucio.cfg"
  - ...
- Added initialization function to:
  - export RUCIO_CFG_ACCOUNT
  - export RUCIO_CFG_RUCIO_HOST
  - export RUCIO_CFG_AUTH_HOST
  - run Jinja2 to create personalised rucio.cfg
  - copy CERN-bundle.pem to shared directory



REANA components

REANA-Job-Controller

- Added Configuration:
  - RUCIO_CONTAINER_NAME = "reana-auth-rucio"
  - RUCIO_CACHE_LOCATION = "/rucio_cache/"
  - RUCIO_CFG_CACHE_FILENAME = "rucio.cfg"
  - ...
- Added initialization function to:
  - export RUCIO_CFG_ACCOUNT
  - export RUCIO_CFG_RUCIO_HOST
  - export RUCIO_CFG_AUTH_HOST
  - run Jinja2 to create personalised rucio.cfg
  - copy CERN-bundle.pem to shared directory



rucio: integrate sidecar container #372

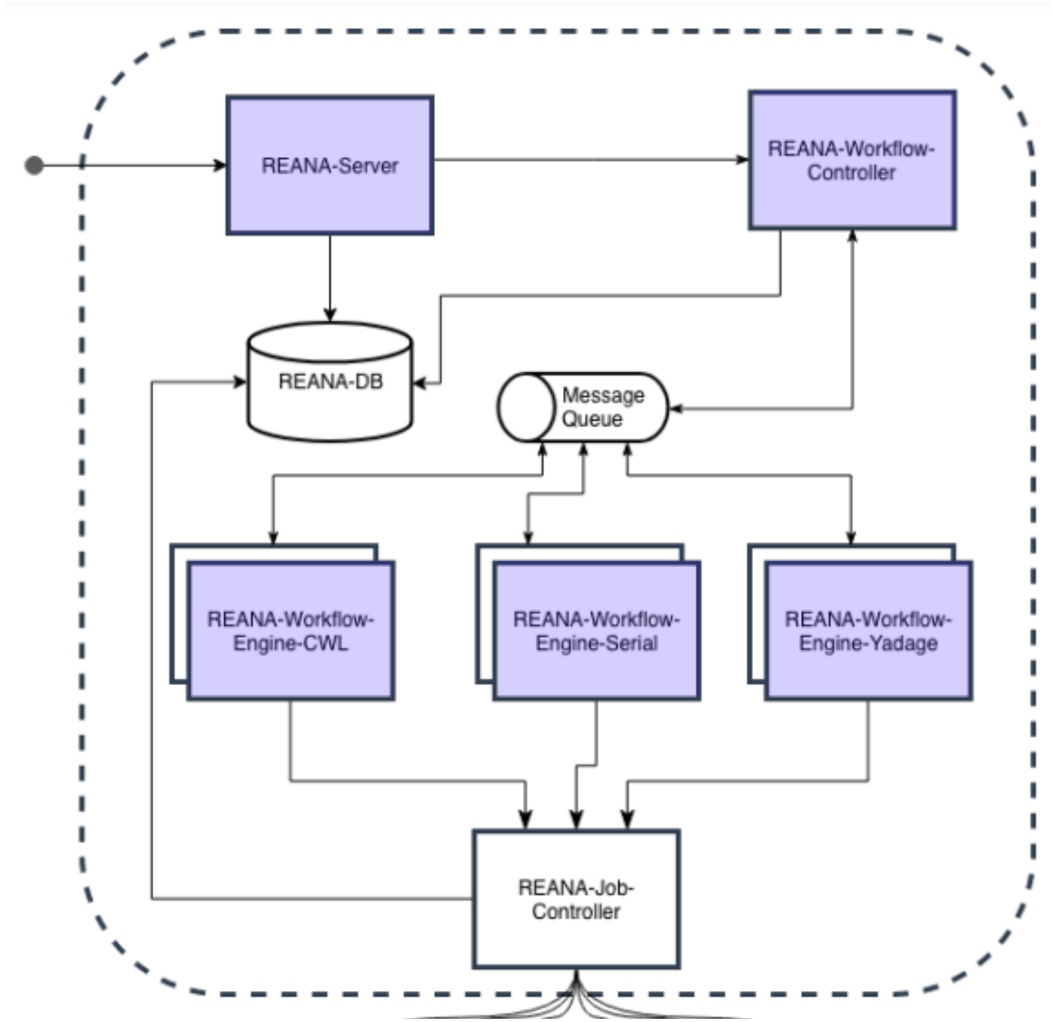
Open kounelisagis wants to merge 2 commits into `reanahub:master` from `kounelisagis:rucio`

REANA components

REANA-Commons
REANA Workflow Engines

Information Propagation:

```
"rucio": {
  "type": "boolean"
},
```

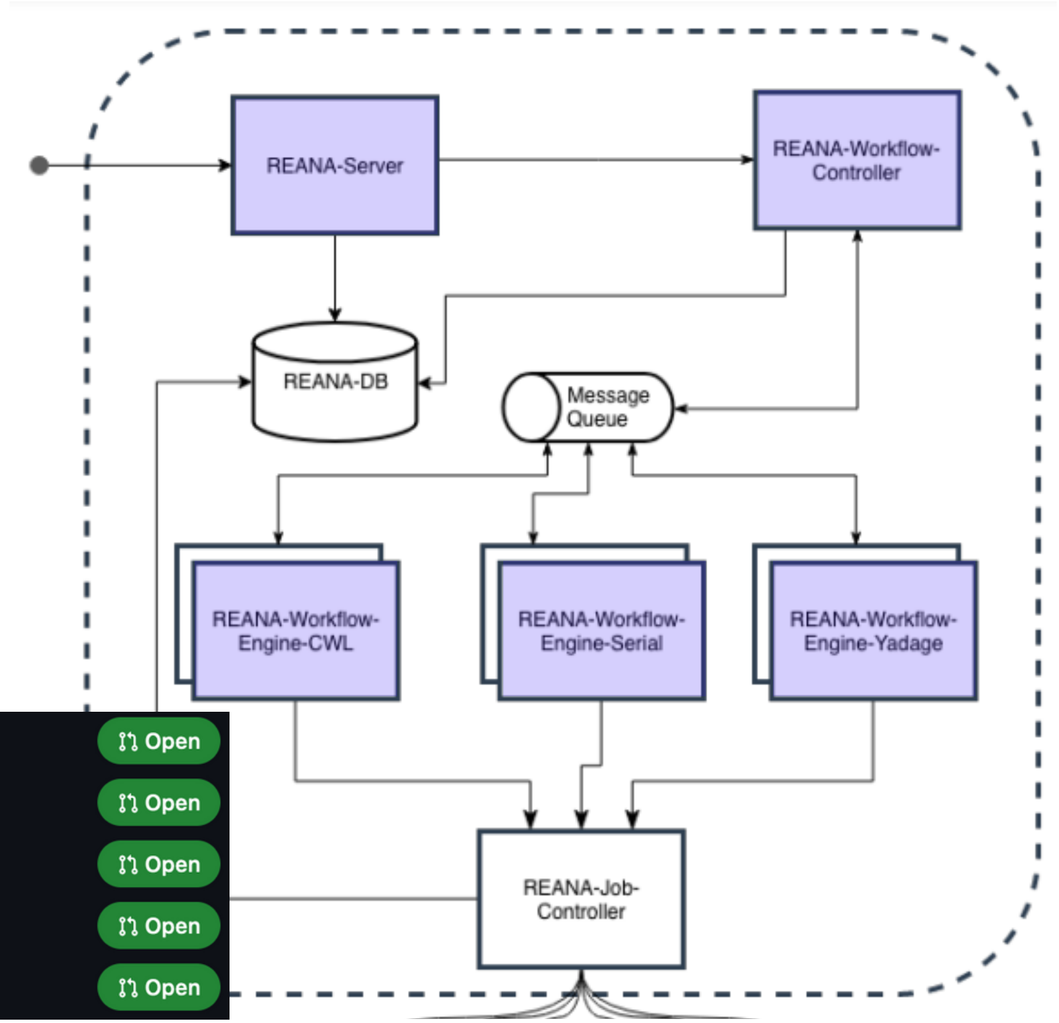


REANA components

REANA-Commons
REANA Workflow Engines

Information Propagation:

```
"rucio": {
  "type": "boolean"
},
```



serial: add field rucio to schema reanahub/reana-commons#366	Open
cwl_reana: add rucio support reanahub/reana-workflow-engine-cwl#233	Open
serial: add rucio support reanahub/reana-workflow-engine-serial#170	Open
executor: add rucio support reanahub/reana-workflow-engine-snakemake#48	Open
yadage: add rucio support reanahub/reana-workflow-engine-yadage#234	Open

Contributions

Rucio Sidecar container	REANA components
Docker environments	Documentation

Docker environments

- Each step need definition of the environment
- In order to run Rucio commands, an environment with rucio-clients installed is required

```
- name: fetchdata
  voms_proxy: true
  rucio: true
  environment: 'projectescape/rucio-client'
  commands:
  - rucio get agis_test:fitdata.C agis_test:gendata.C
```

Docker environments

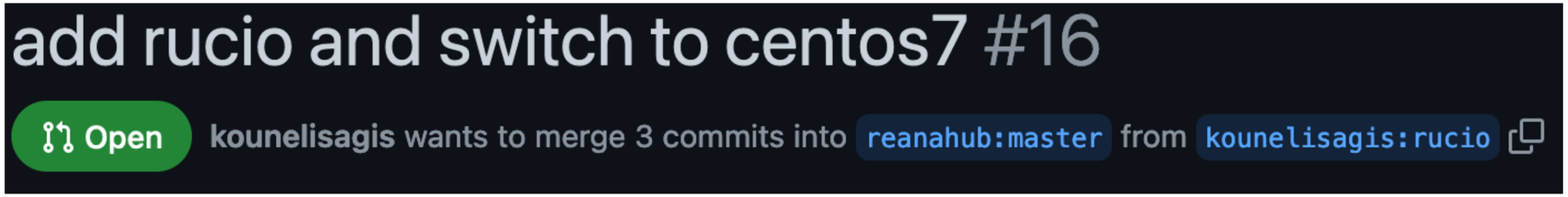
Available Options:

- Use custom environments that support Rucio commands
- Use pre-configured environments for workflows that require Rucio commands:
  - Use projectescape/rucio-client, a basic image with support for Rucio commands
  - **Sidecar container as an environment**
  - Use **reanahub/reana-env-root6**, encapsulated runtime execution environment for ROOT6 based data analyses

Docker environments

Available Options:

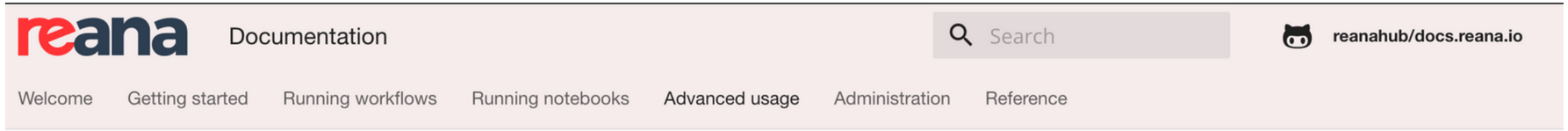
- Use custom environments that support Rucio commands
- Use pre-configured environments for workflows that require Rucio commands:
  - Use projectescape/rucio-client, a basic image with support for Rucio commands
  - **Sidecar container as an environment**
  - Use **reanahub/reana-env-root6**, encapsulated runtime execution environment for ROOT6 based data analyses



Contributions

Rucio Sidecar container	REANA components
Docker environments	Documentation

Documentation



Advanced usage
- Advanced usage
- Compute backends
- Storage backends
- Code repositories
- Access control
  - Access control
  - Kerberos
  - Private Docker registries
  - Rucio
  - VOMS proxy
- Containers
- User quotas

Rucio

About Rucio

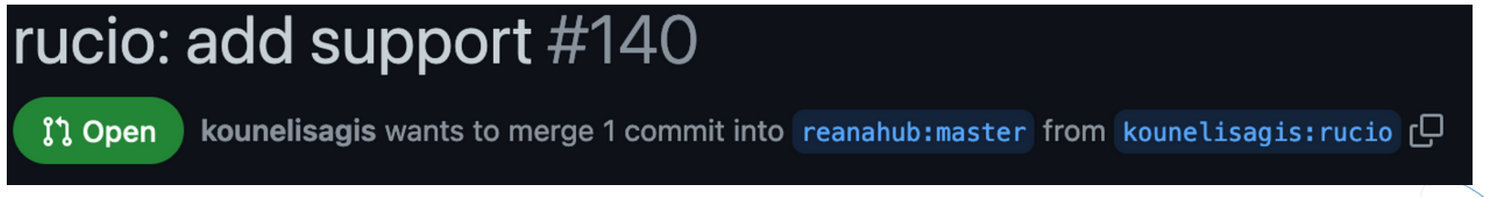
Rucio is a project that provides services and associated libraries for allowing scientific collaborations to manage large volumes of data spread across facilities at multiple institutions and organisations. Rucio was originally developed to meet the requirements of the high-energy physics experiment ATLAS, and now is continuously extended to support the LHC experiments and other diverse scientific communities.

Rucio offers advanced features, is highly scalable, and modular. It is a data management solution that covers the needs of different communities in the scientific domain (e.g., HEP, astronomy, biology).

If your workflow needs data access to a Rucio instance, you can use the Rucio authentication technique with your REANA workflows as described in detail below.

Table of contents

- About Rucio
- VOMS proxy dependency
- Uploading secrets
- Configuring your workflows
- Creating your job environment images



DEMO time!

```yaml
workflow:
  type: serial
  specification:
    steps:
      - name: fetchdata
        voms_proxy: true
        rucio: true
        environment: 'projectescape/rucio-client'
        commands:
        - rucio get agis_test:fitdata.C agis_test:gendata.C
      - name: gendata
        environment: 'reanahub/reana-env-root6:latest'
        kubernetes_memory_limit: '256Mi'
        commands:
        - mkdir -p results && root -b -q 'agis_test/gendata.C(${events},"${data}")'
      - name: fitdata
        environment: 'reanahub/reana-env-root6:latest'
        kubernetes_memory_limit: '256Mi'
        commands:
        - root -b -q 'agis_test/fitdata.C("${data}","${plot}")'
      - name: uploaddata
        voms_proxy: true
        rucio: true
        environment: 'projectescape/rucio-client'
        kubernetes_memory_limit: '256Mi'
        commands:
        - rucio upload --scope agis_test results/plot.png --rse EULAKE-1
```

DEMO time!

ENV variables

```
[root@agisvm-large secrets]# export REANA_SERVER_URL=https://localhost:30443
[root@agisvm-large secrets]# export REANA_ACCESS_TOKEN=[redacted]
```

REANA Secretes

```
(reana) [root@agisvm-large secrets]# reana-client secrets-add --file userkey.pem --file usercert.pem --env VOMSPRO
XY_PASS=[redacted] --env VONAME=escape --env RUCIO_USERNAME=akouneli --overwrite
==> SUCCESS: Secrets VOMSPROXY_PASS, VONAME, RUCIO_USERNAME, userkey.pem, usercert.pem were successfully uploaded.
```

Analysis: create, upload start

```
(reana) [root@agisvm-large reana-demo-root6-roofit]# ./create_upload_start.sh
==> Verifying REANA specification file... /root/reana-demo-root6-roofit/reana.yaml
  -> SUCCESS: Valid REANA specification file.
==> Verifying REANA specification parameters...
  -> SUCCESS: REANA specification parameters appear valid.
==> Verifying workflow parameters and commands...
  -> SUCCESS: Workflow parameters and commands appear valid.
==> Verifying dangerous workflow operations...
  -> SUCCESS: Workflow operations appear valid.
==> Verifying compute backends in REANA specification file...
  -> SUCCESS: Workflow compute backends appear to be valid.
myanalysis.106
==> Detected .gitignore file. Some files might get ignored.
==> SUCCESS: myanalysis is pending
```

DEMO time!

myanalysis #106
Started a few seconds ago

running for 16 seconds
step 1/4

•

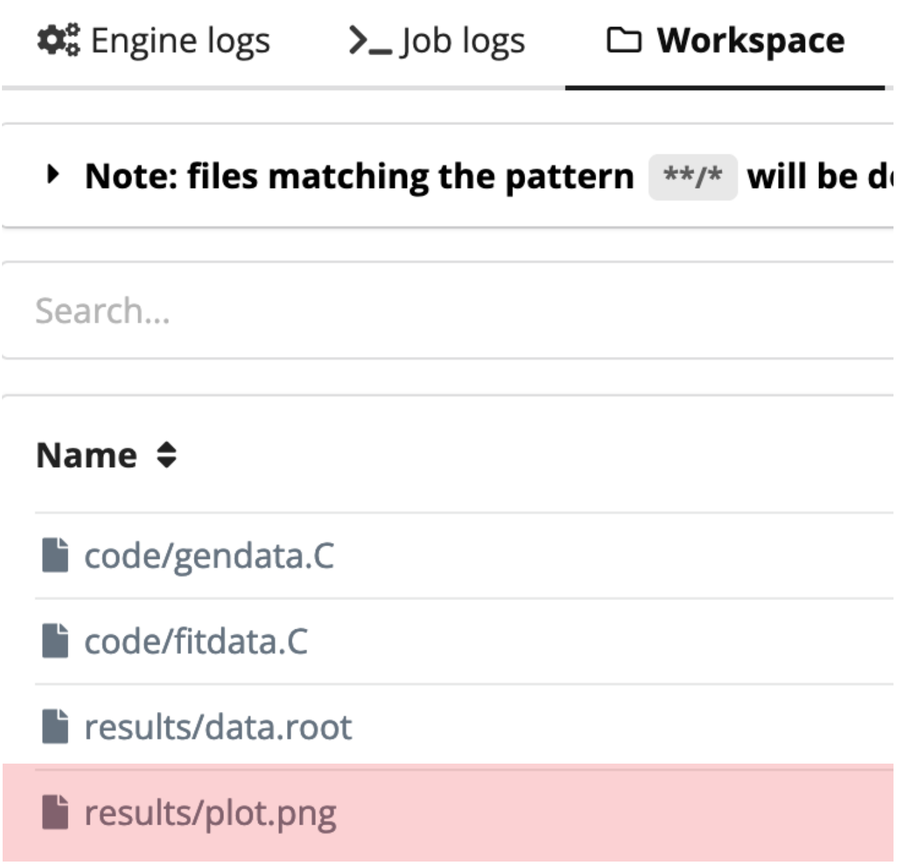
•

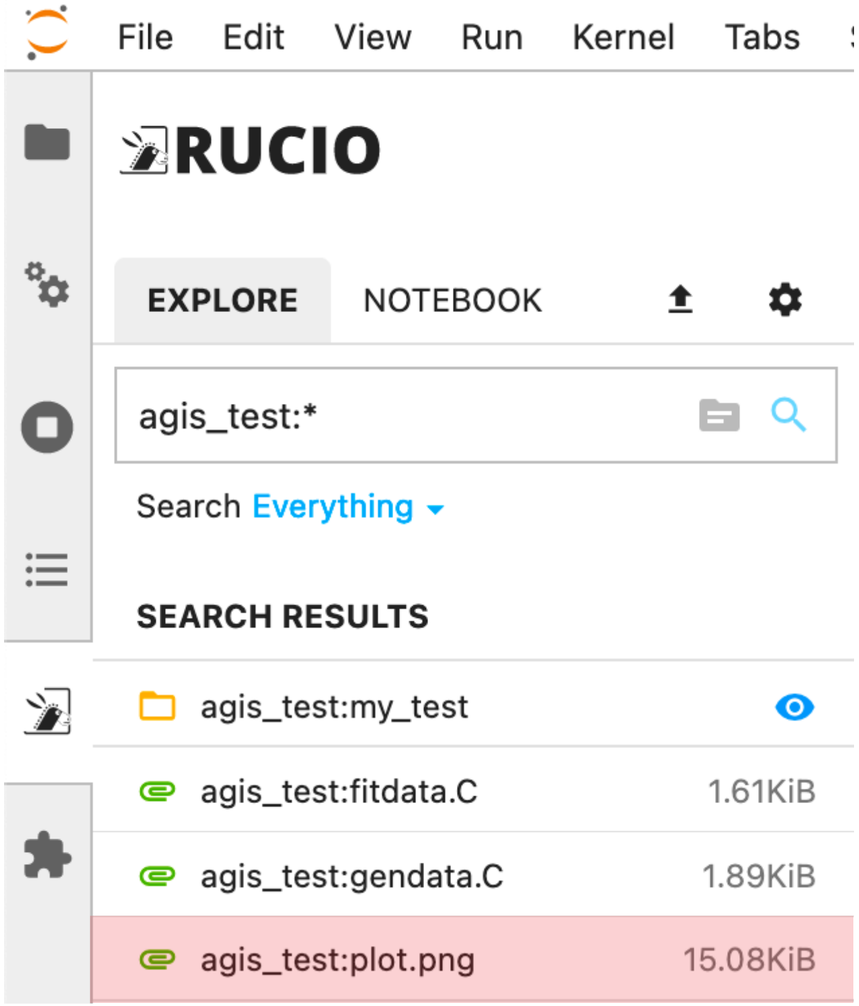
•

myanalysis #106
Finished a few seconds ago

finished in 43 seconds
step 4/4

DEMO time!


Engine logs
Job logs
Workspace
Note: files matching the pattern **/* will be d
Search...
Name
code/gendata.C
code/fitdata.C
results/data.root
results/plot.png


File
Edit
View
Run
Kernel
Tabs
RUCIO
EXPLORE
NOTEBOOK
agis_test:*
Search Everything
SEARCH RESULTS
agis_test:my_test
agis_test:fitdata.C 1.61KiB
agis_test:gendata.C 1.89KiB
agis_test:plot.png 15.08KiB

Further implementations

Provide abstraction of inputs field: the stage-in and stage-out processes from external sources, such as Rucio, could be automatised.

↓

No need for direct rucio client calls.

↓

Enrich reana.yaml input files section to support a syntax like
rucio(my_instance, my_scope, my_fileID)

THANK YOU!