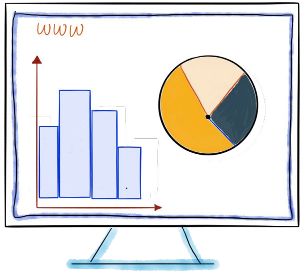
ESCAPE: a review of idea for Data + Analysis challenge using ATLAS Open Data



Arturo Sánchez Pineda - LAPP

26th (from 23rd) April, 2021

Overview

This is an attempt to describe a series of exercises to be performed during the ESCAPE Data and Analysis Challenge the next November

- Data “multiplication” where multiple version of the same data is generated, simulating a data-augmentation process
    - Requesting data from the Datalake at higher rate than the analysis
- Writing of such “multiplied” data back to the Datalake
    - Defining different RSEs
- Exercises include the analysis of data stored in the Datalake
    - Writing back the results into the Datalake (small files of ~100’s kb size each)
    - Analysis can be perform using CLI or the JupyterLab UI
- Create clear instructions for users/computers that can be part of the challenge

Data Multiplication

Artificial multiplication of the data

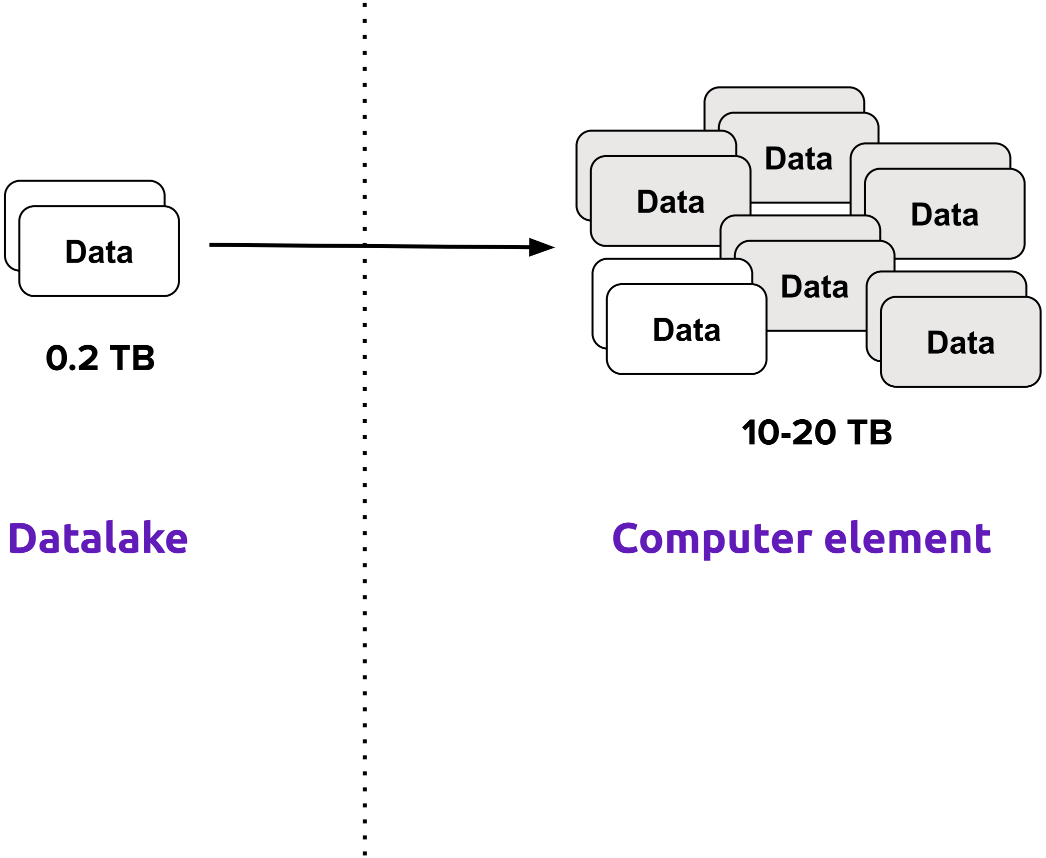
ROOT files can be added when they share an internal structure (i.e. same trees inside). Called “hadd”

We can profit from that property to artificially multiply the datasets.

This process allows augmenting the data to any arbitrary value.

We can use that augmented data to run the analysis examples

- Of course, the results are meaningless.



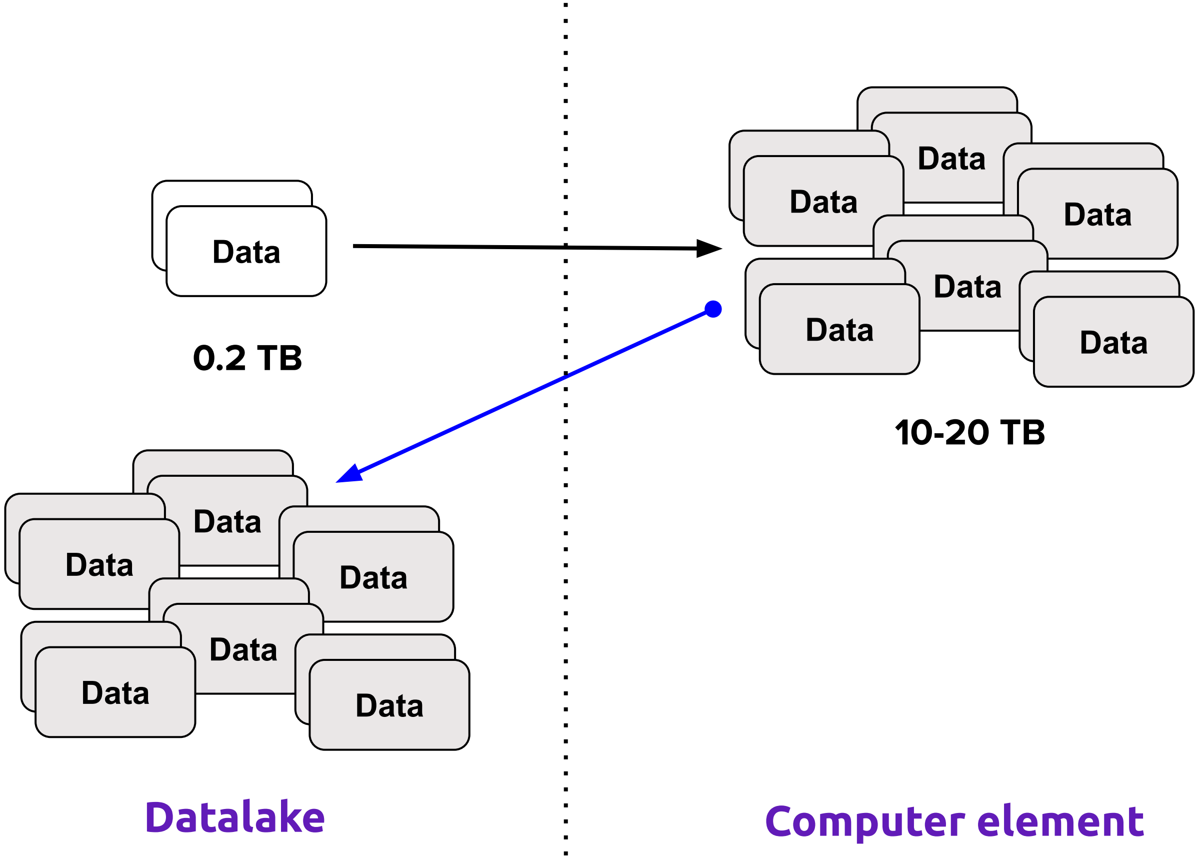
Artificial multiplication of the data

The “augmented” data

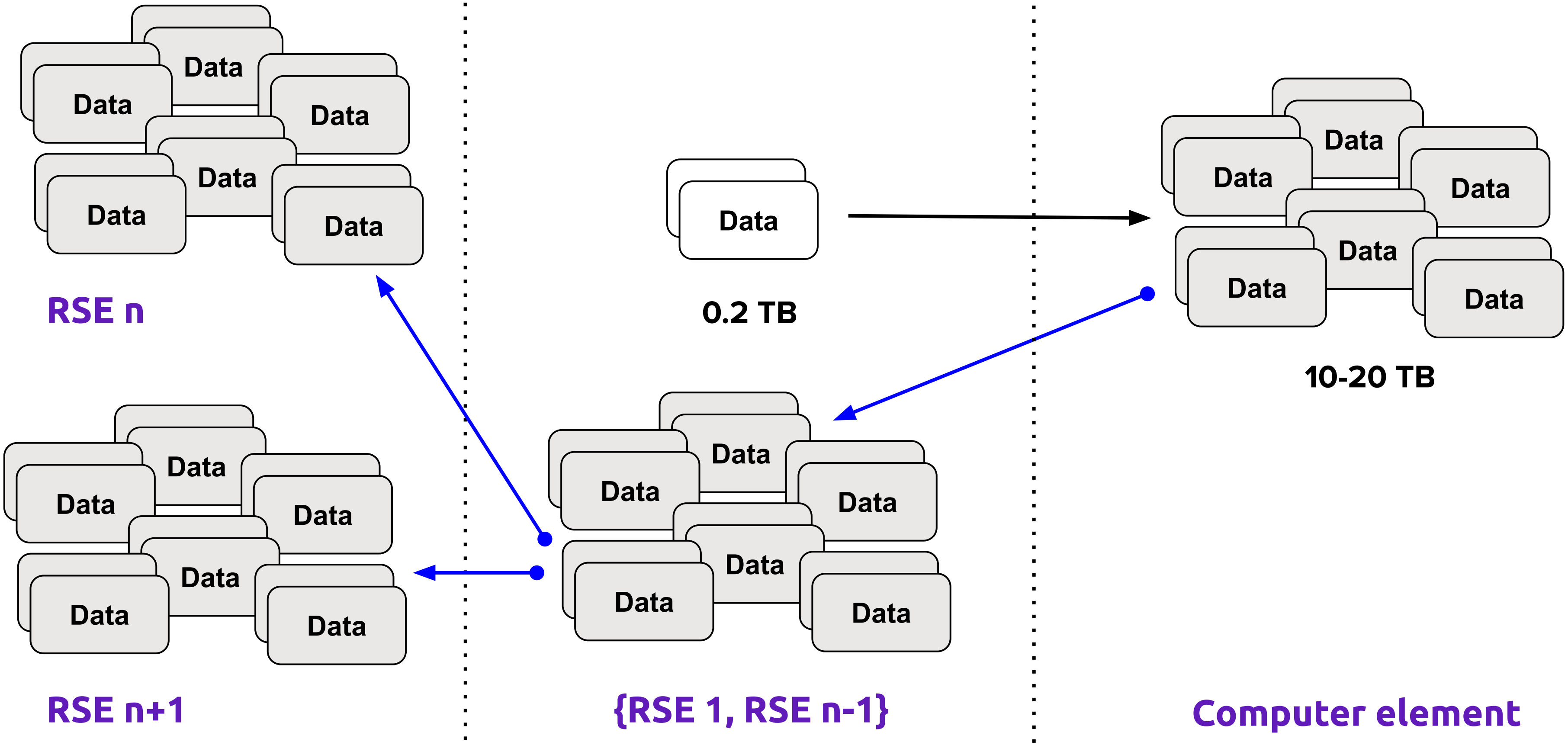
- More and/or larger files
- The process to add multiples files also use computing

After (or during the process), the code automatically can write such new data back to the Datalake

- We can also replicate in multiples RSEs as part of the challenge’s tests



Artificial multiplication of the data



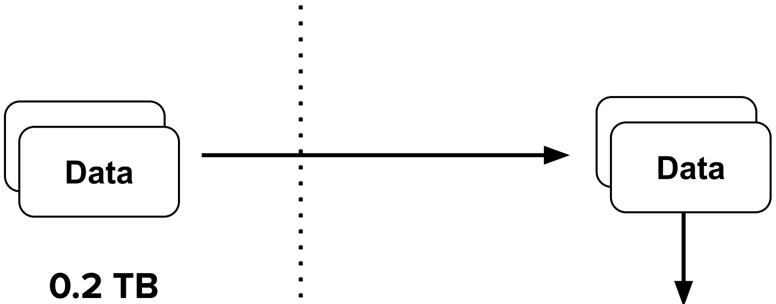
Data Analysis

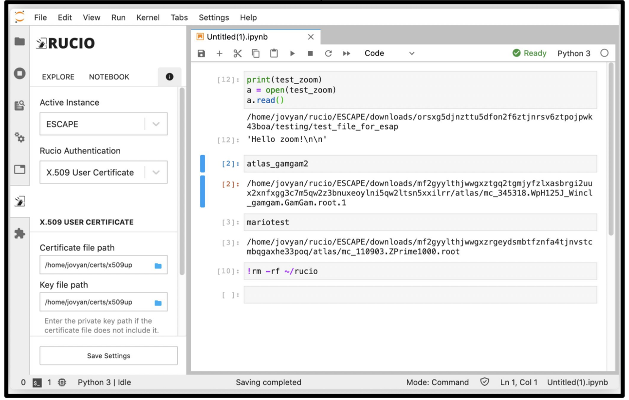
Analysis examples

We can use the current ATLAS Open Data analysis examples to retrieve and use datasets from the Datalake

- Analysis can be notebooks or analysis frameworks
- They can take from a few minutes (e.g. 5-30 min)
- To several (e.g. 4) hours

Also, write back the outputs





Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the certificate file does not include it.

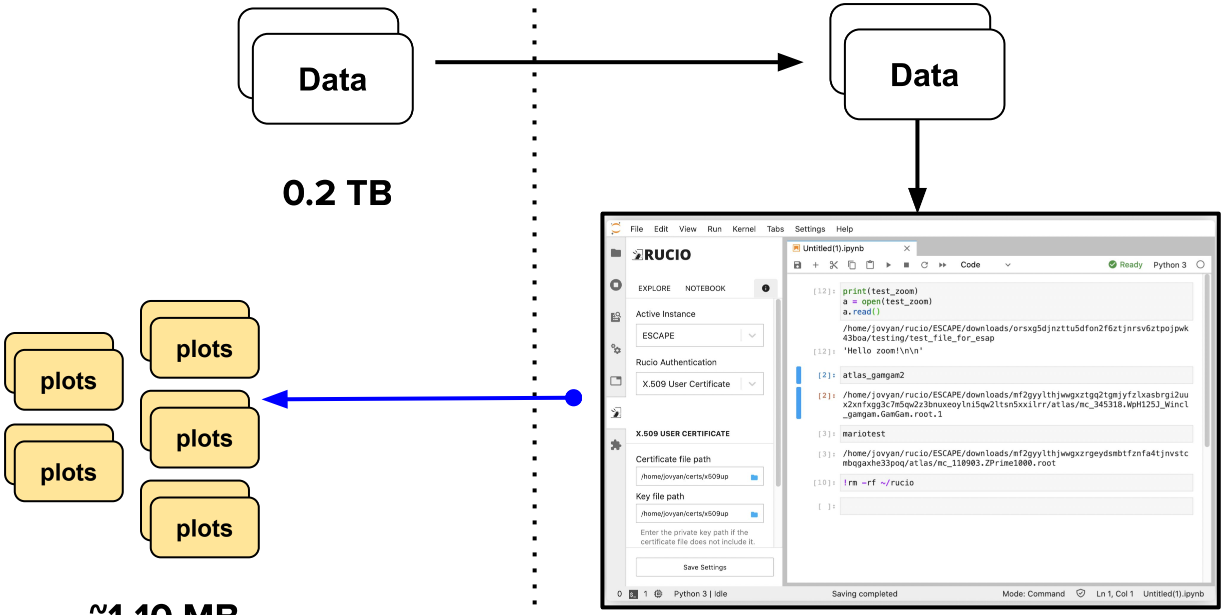
```
[12]: print(test_zoom)
      a = open(test_zoom)
      a.read()

      /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk
      43boa/testing/test_file_for_esap
[12]: 'Hello zoom!\n\n'

[2]: atlas_gamgam2

[2]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uu
     x2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl
     _gamgam.GamGam.root.1

[3]: mariotest

[3]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstc
     mbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root

[10]: !rm -rf ~/rucio

[ ]:
```

0 1 Python 3 | Idle   Saving completed   Mode: Command   Ln 1, Col 1   Untitled(1).ipynb

Analysis examples

The outputs of the analysis can be upload to the Datalake

- The outputs are small; they are plots that can also be store in ROOT files
- No intention to upload single PNG files to The Datalake



0.2 TB

~1-10 MB

Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the certificate file does not include it.

```
[12]: print(test_zoom)
      a = open(test_zoom)
      a.read()

      /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk
      43boa/testing/test_file_for_esap
[12]: 'Hello zoom!\n\n'

 [2]: atlas_gamgam2

 [2]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uu
      x2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl
      _gamgam.GamGam.root.1

 [3]: mariotest

 [3]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstc
      mbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root

[10]: !rm -rf ~/rucio

 [ ]:
```

0 1 Python 3 | Idle   Saving completed   Mode: Command   Ln 1, Col 1   Untitled(1).ipynb

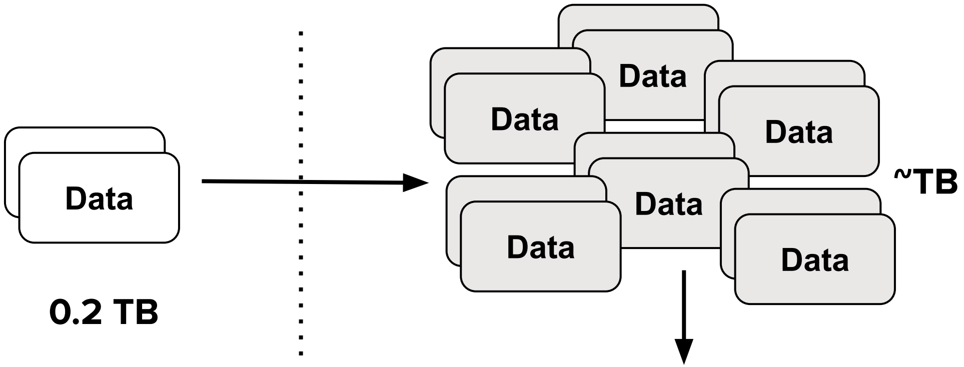
Datalake

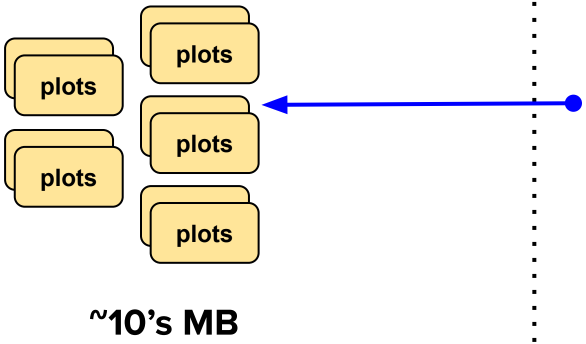
Computer Element

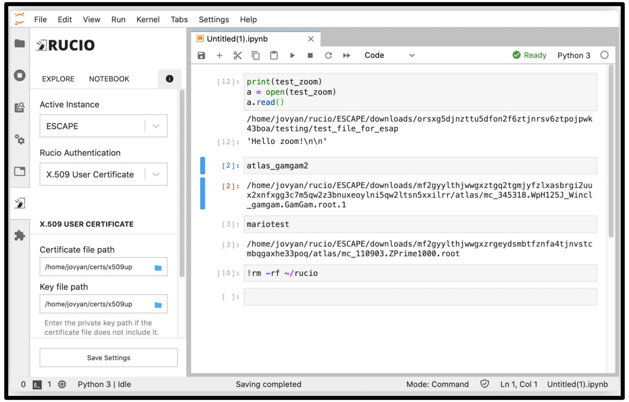
Analysis examples

We can also run the analysis examples over the “multiplied” data

- This can help to simulate longer analysis that can last several hours (e.g. ~8-12 hours)
- In case this kind of “stress” is useful in this challenge







Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

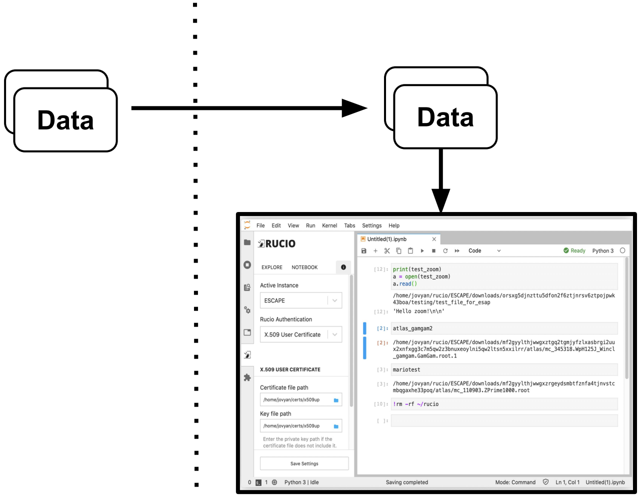
Enter the private key path if the certificate file does not include it.

```
[12]: print(test_zoom)
      a = open(test_zoom)
      a.read()

      /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk
      43boa/testing/test_file_for_esap
[12]: 'Hello zoom!\n\n'

[2]:  atlas_gamgam2

[2]:  /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uu
      x2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl
      _gamgam.GamGam.root.1

[3]:  mariotest

[3]:  /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstc
      mbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root

[10]: !rm -rf ~/rucio

[ ]:
```

Analysis examples

We can use multiple computer elements performing analysis at the same time

- multiple users with a single machine and/or a single user with multiple machines



Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

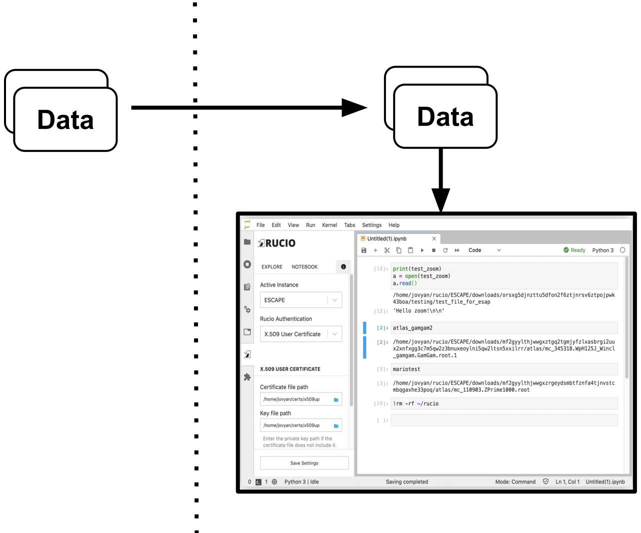
X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the certificate file does not include it.

```
[12]: print(test_zoom)
      a = open(test_zoom)
      a.read()

      /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk
      43boa/testing/test_file_for_esap
[12]: 'Hello zoom!\n\n'

[2]:  atlas_gamgam2

[2]:  /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uu
      x2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl
      _gamgam.GamGam.root.1

[3]:  mariotest

[3]:  /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstc
      mbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root

[10]: !rm -rf ~/rucio

[ ]:
```



Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

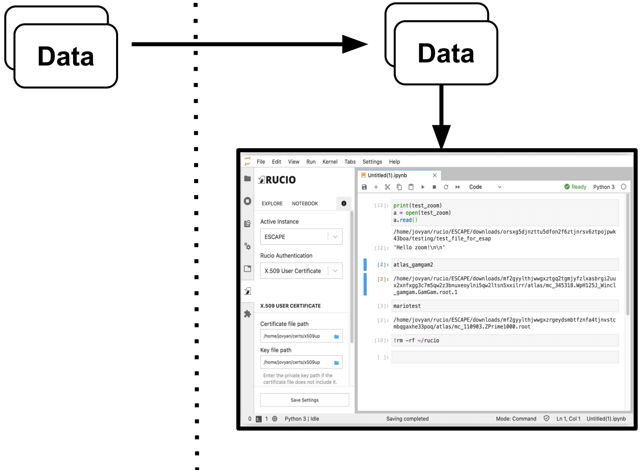
X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the certificate file does not include it.

[12]: print(test_zoom) a = open(test_zoom) a.read() /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk43boa/testing/test_file_for_esap [12]: 'Hello zoom!\n\n' [2]: atlas_gamgam2 [2]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uux2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl_gamgam.GamGam.root.1 [3]: mariotest [3]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstcmbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root [10]: !rm -rf ~/rucio [ ]:



Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the certificate file does not include it.

[12]: print(test_zoom) a = open(test_zoom) a.read() /home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk43boa/testing/test_file_for_esap [12]: 'Hello zoom!\n\n' [2]: atlas_gamgam2 [2]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uux2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl_gamgam.GamGam.root.1 [3]: mariotest [3]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstcmbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root [10]: !rm -rf ~/rucio [ ]:

● ● ●

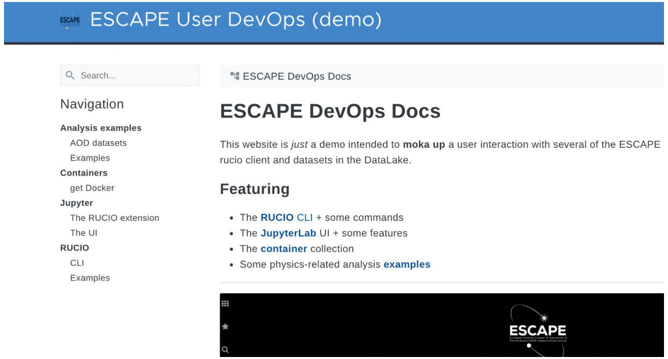
Summary

This set of slides intends to look for feedback on the possible activities to be performed during the next ESCAPE data and analysis challenge

Many of the processes will require the creation of scripts to “automatise” some of those processes

As one of the previous step, I would like to replicate those ~220GB of open data to as many sites as possible (already existing in Fr&It RSEs)

Another essential part of the job will be to prepare instructions so others can replicate the same exercises and simulate a more realistic scenario: multiple users.



ESCAPE DevOps Docs

This website is *just* a demo intended to **moka up** a user interaction with several of the ESCAPE rucio client and datasets in the DataLake.

Featuring

- The **RUCIO** CLI + some commands
- The **JupyterLab** UI + some features
- The **container** collection
- Some physics-related analysis **examples**

Backup

ATLAS Open Data datasets in the Datalake

- ROOT yet need to be importable from a notebook
  - It is deployed for testing in DockerHub
- Add more datasets to the Datalake
  - All the 13 TeV and 8 TeV ATLAS Open Data samples
  - 16 datasets → 940 samples (ROOT files)
  - < 200 GB
  - Scope used: **ATLAS_OD_EDU** (for **ATLAS O**pen **D**ata for **EDU**cation)
  - Source of the datasets:

    http://opendata.atlas.cern/samples-13tev/ **&** http://opendata.atlas.cern/samples-8tev/
  - Another set of 10 ROOT files to come (dedicated Jet MC samples) → 1 dataset, ~21 GB.

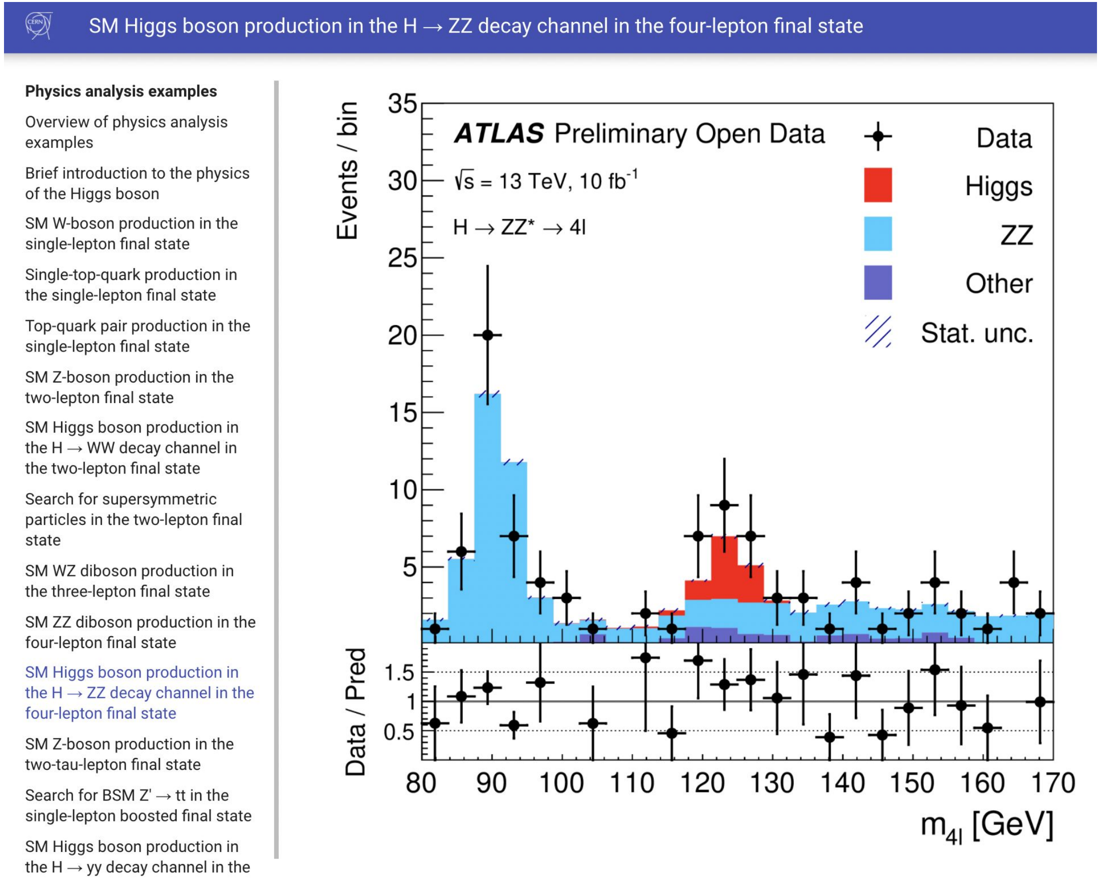
ATLAS Open Data → C++ examples framework

To run C++ analyses

More computational-complex particle physics analysis examples using the existing publicly available data

More in [Opendata.atlas.cern - documentation 13 TeV - physics]()

Also use PROOF, adding a parallel component to the examples.



SM Higgs boson production in the H → ZZ decay channel in the four-lepton final state

Physics analysis examples

Overview of physics analysis examples

Brief introduction to the physics of the Higgs boson

SM W-boson production in the single-lepton final state

Single-top-quark production in the single-lepton final state

Top-quark pair production in the single-lepton final state

SM Z-boson production in the two-lepton final state

SM Higgs boson production in the H → WW decay channel in the two-lepton final state

Search for supersymmetric particles in the two-lepton final state

SM WZ diboson production in the three-lepton final state

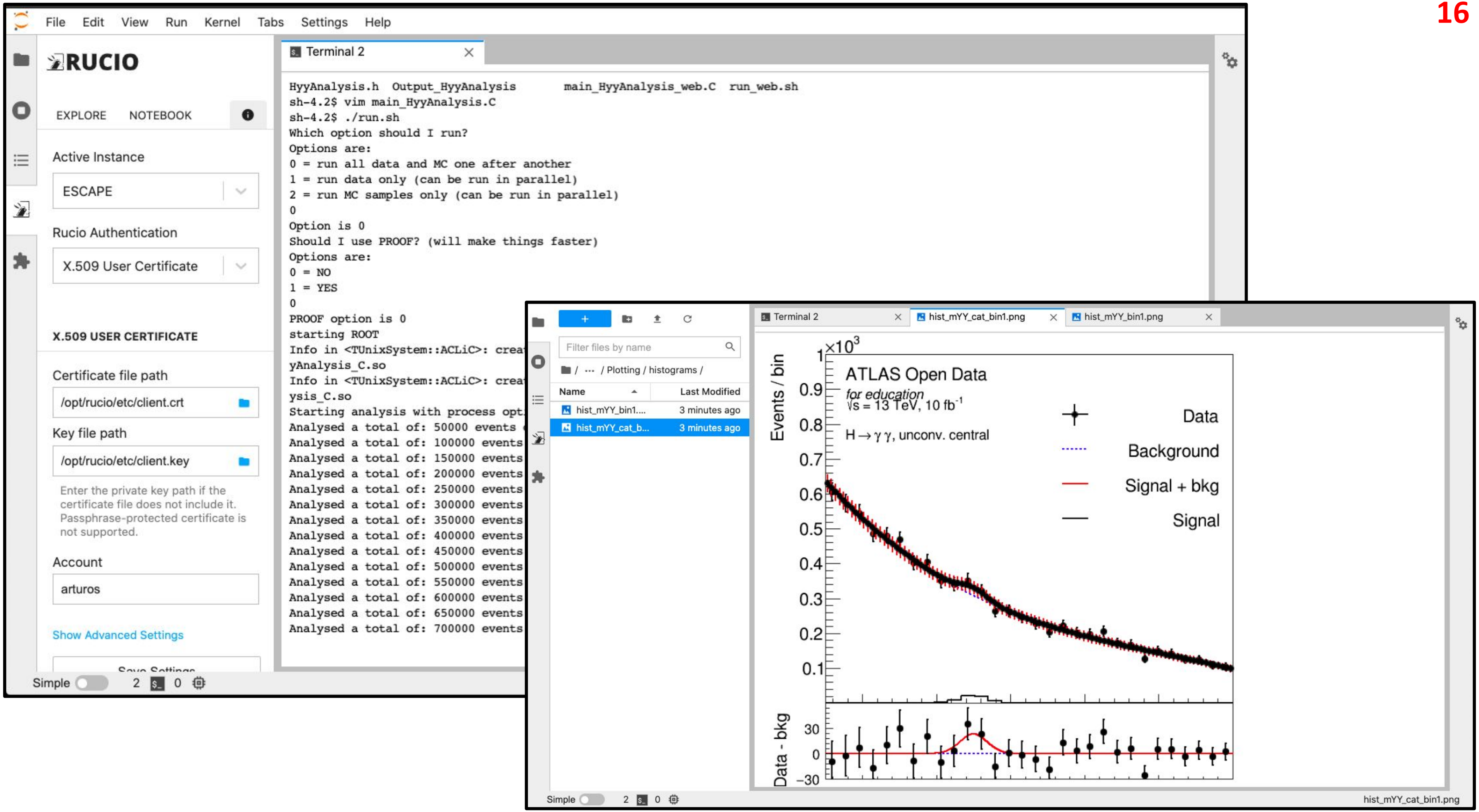
SM ZZ diboson production in the four-lepton final state

SM Higgs boson production in the H → ZZ decay channel in the four-lepton final state

SM Z-boson production in the two-tau-lepton final state

Search for BSM Z' → tt in the single-lepton boosted final state

SM Higgs boson production in the H → yy decay channel in the



Active Instance

ESCAPE

Rucio Authentication

X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path

/opt/rucio/etc/client.crt

Key file path

/opt/rucio/etc/client.key

Enter the private key path if the certificate file does not include it. Passphrase-protected certificate is not supported.

Account

arturos

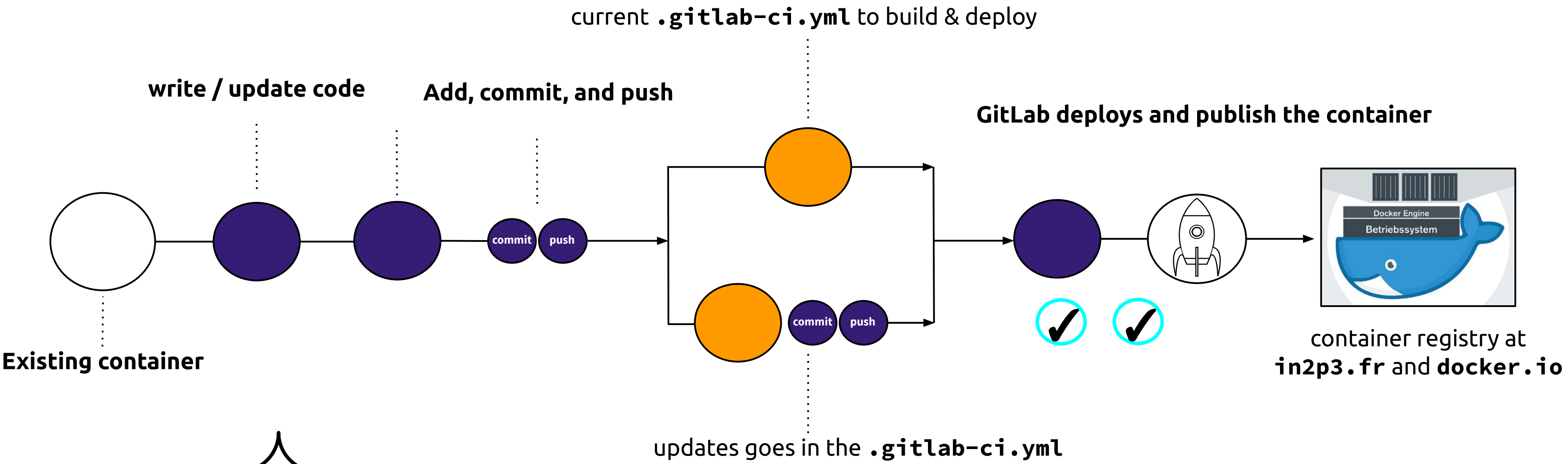
Show Advanced Settings

Save Settings

```
HyyAnalysis.h  Output_HyyAnalysis       main_HyyAnalysis_web.C  run_web.sh
sh-4.2$ vim main_HyyAnalysis.C
sh-4.2$ ./run.sh
Which option should I run?
Options are:
0 = run all data and MC one after another
1 = run data only (can be run in parallel)
2 = run MC samples only (can be run in parallel)
0
Option is 0
Should I use PROOF? (will make things faster)
Options are:
0 = NO
1 = YES
0
PROOF option is 0
starting ROOT
Info in <TUnixSystem::ACLiC>: crea
yAnalysis_C.so
Info in <TUnixSystem::ACLiC>: crea
ysis_C.so
Starting analysis with process opt
Analysed a total of: 50000 events
Analysed a total of: 100000 events
Analysed a total of: 150000 events
Analysed a total of: 200000 events
Analysed a total of: 250000 events
Analysed a total of: 300000 events
Analysed a total of: 350000 events
Analysed a total of: 400000 events
Analysed a total of: 450000 events
Analysed a total of: 500000 events
Analysed a total of: 550000 events
Analysed a total of: 600000 events
Analysed a total of: 650000 events
Analysed a total of: 700000 events
```

Name	Last Modified
hist_mYY_bin1....	3 minutes ago
hist_mYY_cat_b...	3 minutes ago

A view to the current container



current **`.gitlab-ci.yml`** to build & deploy

write / update code

Add, commit, and push

GitLab deploys and publish the container

Existing container

container registry at
`in2p3.fr` and **`docker.io`**

updates goes in the **`.gitlab-ci.yml`**

Container CI / CD

- The series of resources is package in a single container
- The CI setup automatically handles the publication of the container

Frédéric &
Berkay's job

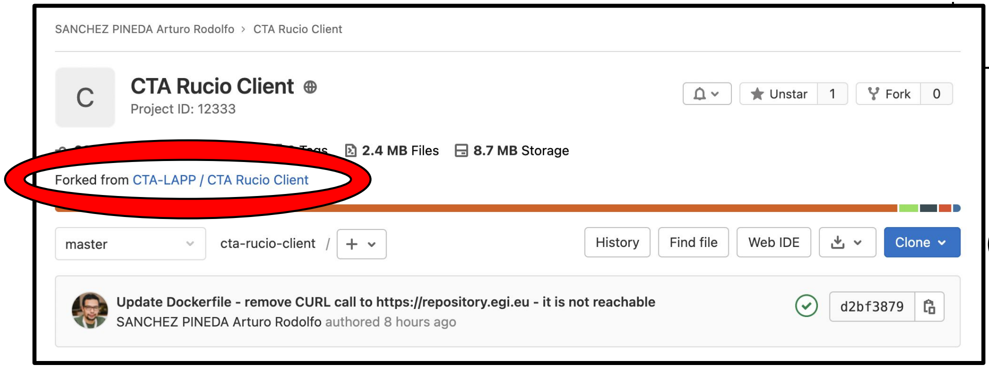
Several developments and deployment already in place

- The compendium of resources includes the current rucio client + JupyterLab + RUCIO extension, proxy & authentication, ...

A view to the current container

current **`.gitlab-ci.yml`** to build & deploy

write / update code **Add, commit, and push**



SANCHEZ PINEDA Arturo Rodolfo › CTA Rucio Client

CTA Rucio Client
Project ID: 12333

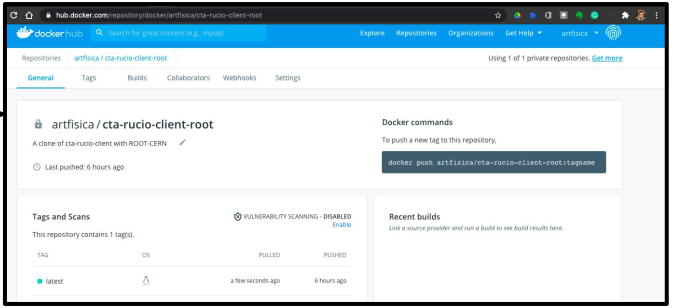
2.4 MB Files 8.7 MB Storage

Forked from CTA-LAPP / CTA Rucio Client

Update Dockerfile - remove CURL call to https://repository.egi.eu - it is not reachable
SANCHEZ PINEDA Arturo Rodolfo authored 8 hours ago

d2bf3879

GitLab deploys and publish the container



hub.docker.com/repository/docker/artfisica/cta-rucio-client-root

Using 1 of 1 private repositories. Get more

artfisica / cta-rucio-client-root

A clone of cta-rucio-client with ROOT-CERN

Last pushed: 6 hours ago

Docker commands

To push a new tag to this repository,

```
docker push artfisica/cta-rucio-client-root:tagname
```

Tags and Scans

VULNERABILITY SCANNING - DISABLED
Enable

This repository contains 1 tag(s).

TAG	OS	PULLED	PUSHED
latest		a few seconds ago	6 hours ago

Recent builds

Link a source provider and run a build to see build results here.

updates goes in the **`.gitlab-ci.yml`**

Container CI / CD

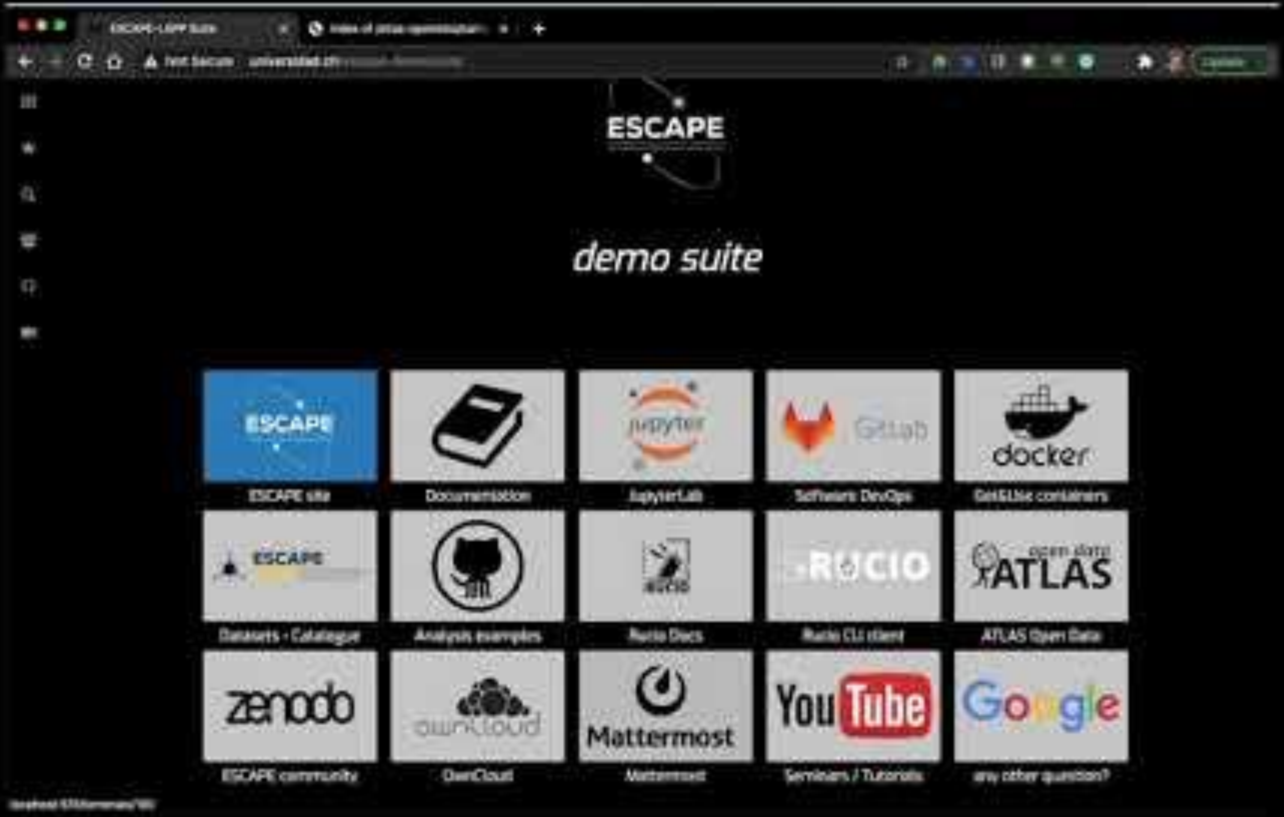
- The series of resources is package in a single container
- The CI setup automatically handles the publication of the container

Several tools and updates added

- Mainly ROOT + some dependencies and extra tools…
- Jupyter conf file to handle the usage of the rucio extension (Muhammad feedback, see later)
- From JupyterLab-3 the widgets are installed using ipywidgets instead of labextension

The RUCIO CLI client

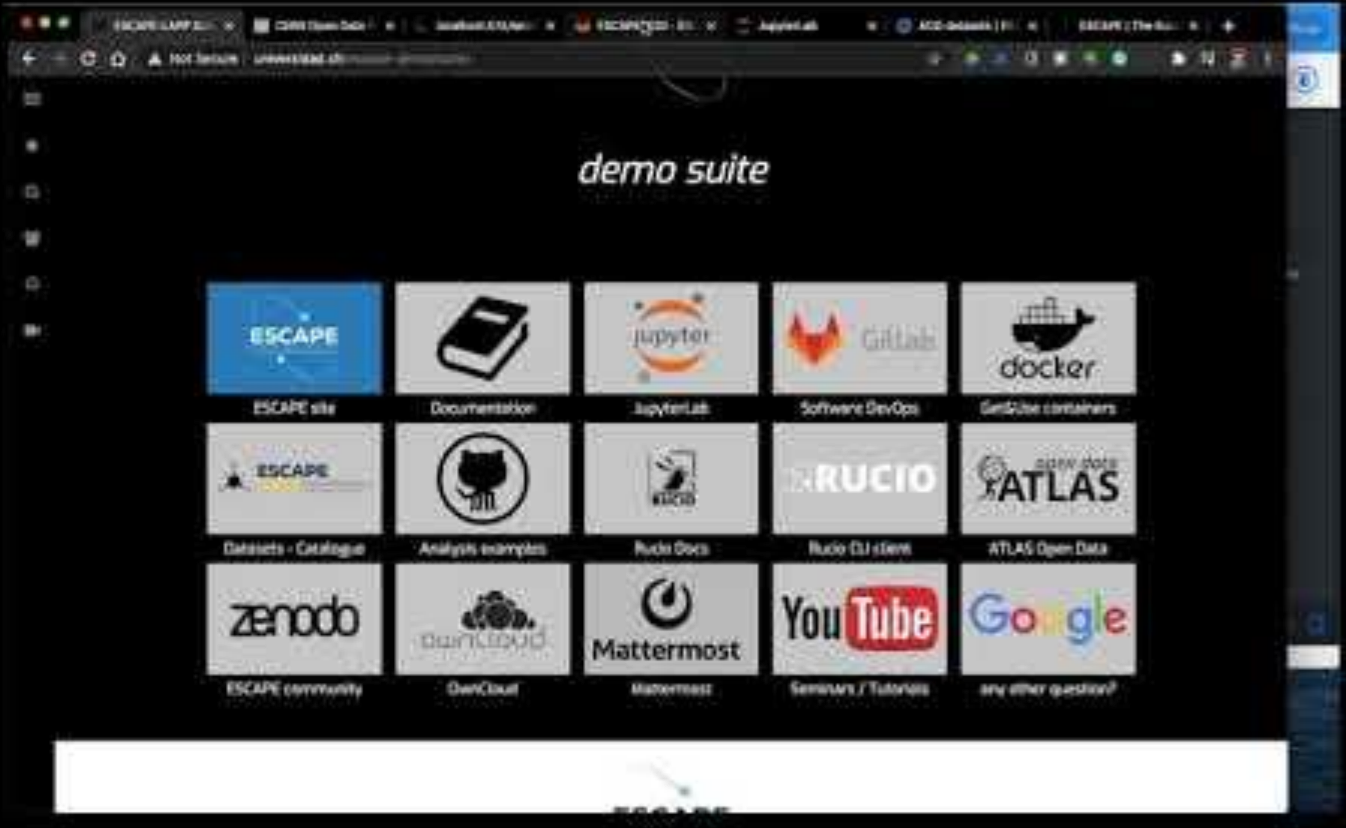
(a 90 sec video, mainly for new users)



http://universidad.ch/escape-demo/suite/

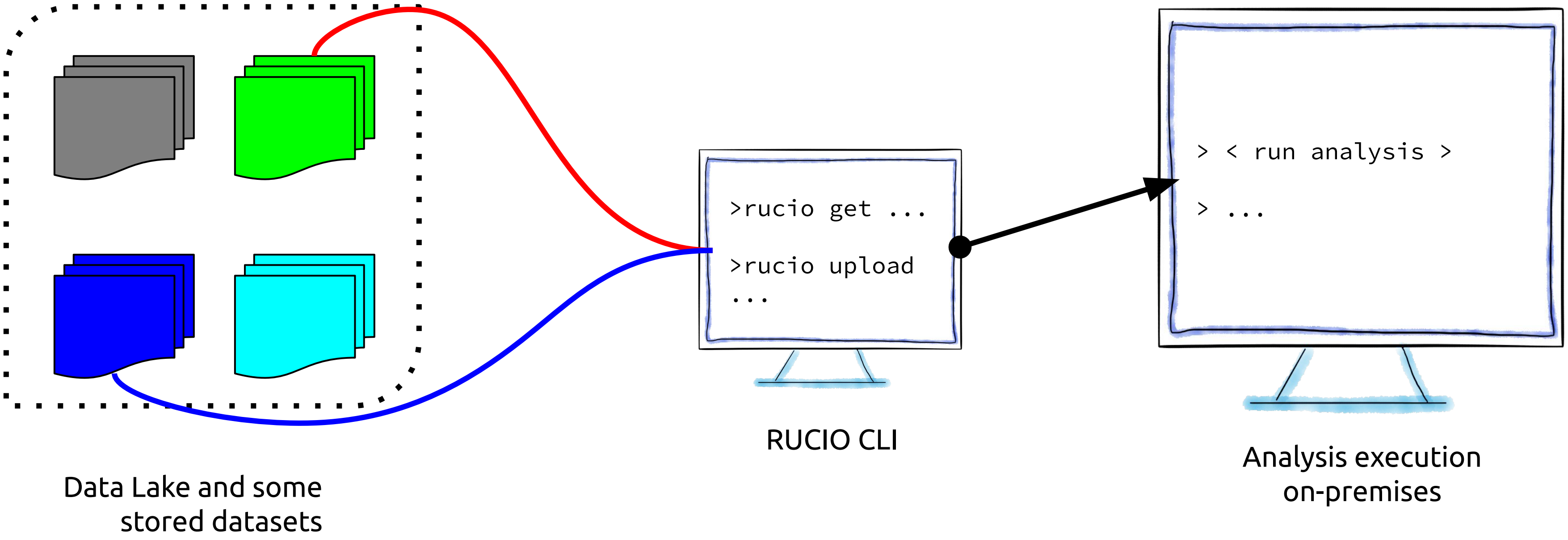
Ongoing developments with **JupyterLab & RUCIO extension**

(a 150 sec video)



More tools to finish to integrate in the container, like more kernels, PROOF, CVMFS

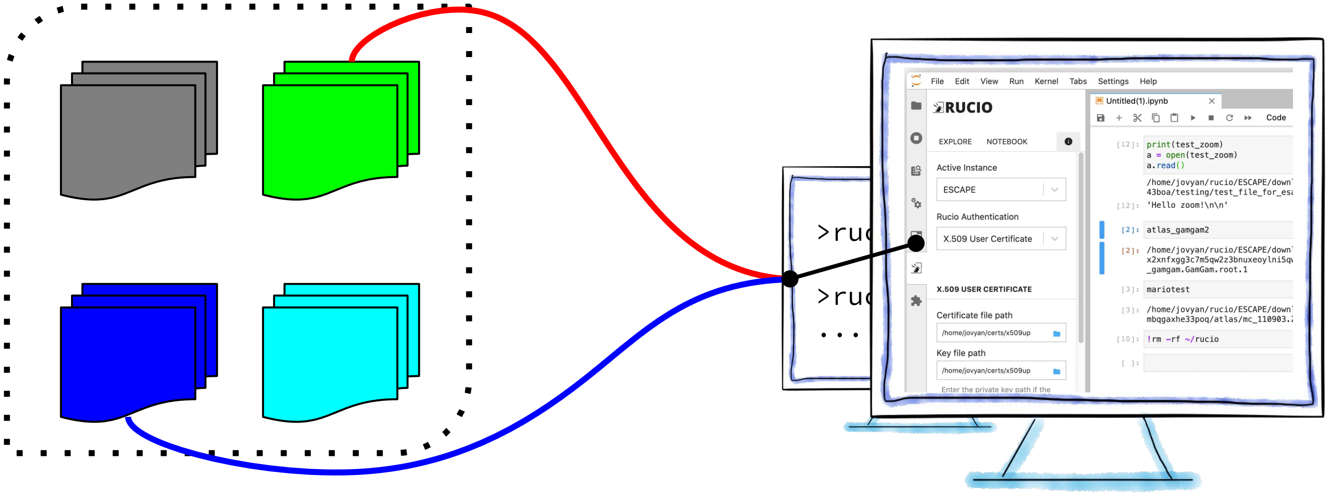
http://universidad.ch/escape-demo/suite/



```
>rucio get ...

>rucio upload
...
```

```
> < run analysis >

> ...
```

CLI interaction with samples



Active Instance: ESCAPE

Rucio Authentication: X.509 User Certificate

X.509 USER CERTIFICATE

Certificate file path: /home/jovyan/certs/x509up

Key file path: /home/jovyan/certs/x509up

Enter the private key path if the

```
[12]: print(test_zoom)
      a = open(test_zoom)
      a.read()

      /home/jovyan/rucio/ESCAPE/down
      43boa/testing/test_file_for_es
[12]: 'Hello zoom!\n\n'

[2]:  atlas_gamgam2

[2]:  /home/jovyan/rucio/ESCAPE/down
      x2xnfxgg3c7m5qw2z3bnuxeoylni5q
      _gamgam.GamGam.root.1

[3]:  mariotest

[3]:  /home/jovyan/rucio/ESCAPE/down
      mbqgaxhe33poq/atlas/mc_110903.

[10]: !rm -rf ~/rucio

[ ]:
```

Authentication, discovery and download of the samples using the Web UI

Data Lake and stored Open Data datasets

Container with RUCIO + JupyterLab extension

RUCIO+JupyterLab (container) interaction for users