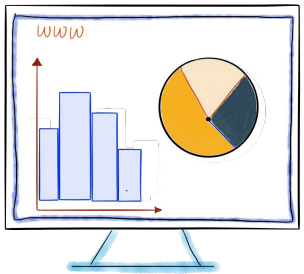
ESCAPE: a view of RUCIO + JupyterLab + ATLAS Open Data integration



Arturo Sánchez Pineda

22 January 2021, LAPP (virtual)

Overview

This is a first attempt to summarise the activities relative to *a* integration and consolidation of the Data Lake via RUCIO and a friendly web-based UI like JupyterLab.

And how ATLAS Open Data is used as a Test for such technology.

Caveats:
Since I am new to the project and team, I may not have the proper jargon of clear concepts, yet. Any feedback is very welcome :)

* Stopped to pursue a PhD in particle physics.

User’s context

The user’s context

In this case, the target audience refers to scientists & advanced users looking for data to perform or reproduce an analysis.

They are/should be aware of the RUCIO as a service, but enjoying the UI and features of a tool as JupyterLab.

And because this is a tool intrinsically web-based, it can be used in a cloud computing environment. So, in terms of setup, it points also to the institute sysadmins that set those tools for their academic community.

RUCIO & JupyterLab

The JupyterLab RUCIO plugin

In 2020 at CERN, Muhammad Aditya Hilmy created a JupyterLab extension that allows the proper authentication (login/pass or certificate) and access to the datasets in the Data Lake using RUCIO.

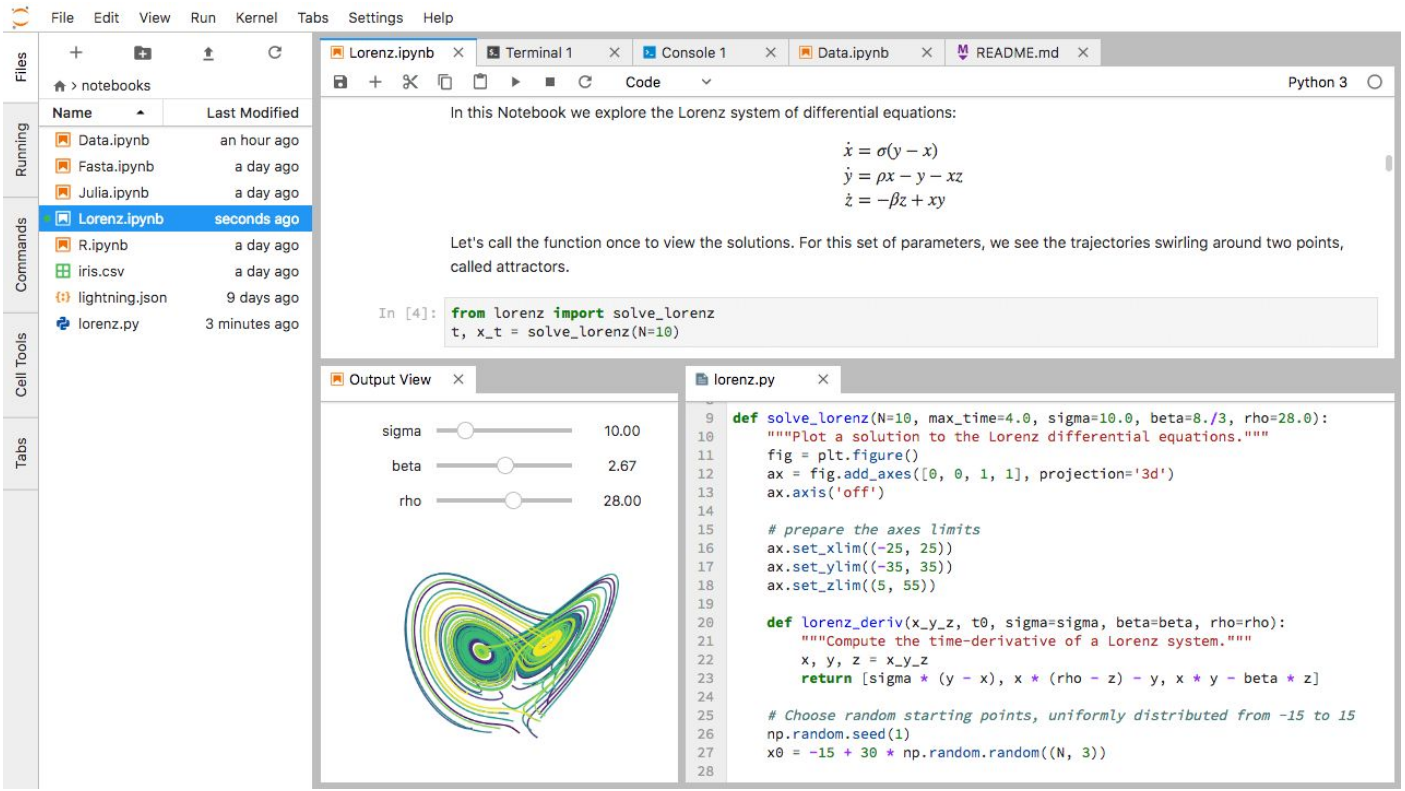
More on how it looks like in one of Muhammad's presentations

The main idea is to deliver an easy and transparent way to access, download and use datasets replicated in the Data Lake.
It hides all the complexity on that access and allows a seamless usage of the data in a Jupyter notebook analysis.

The traditional JupyterLab UI

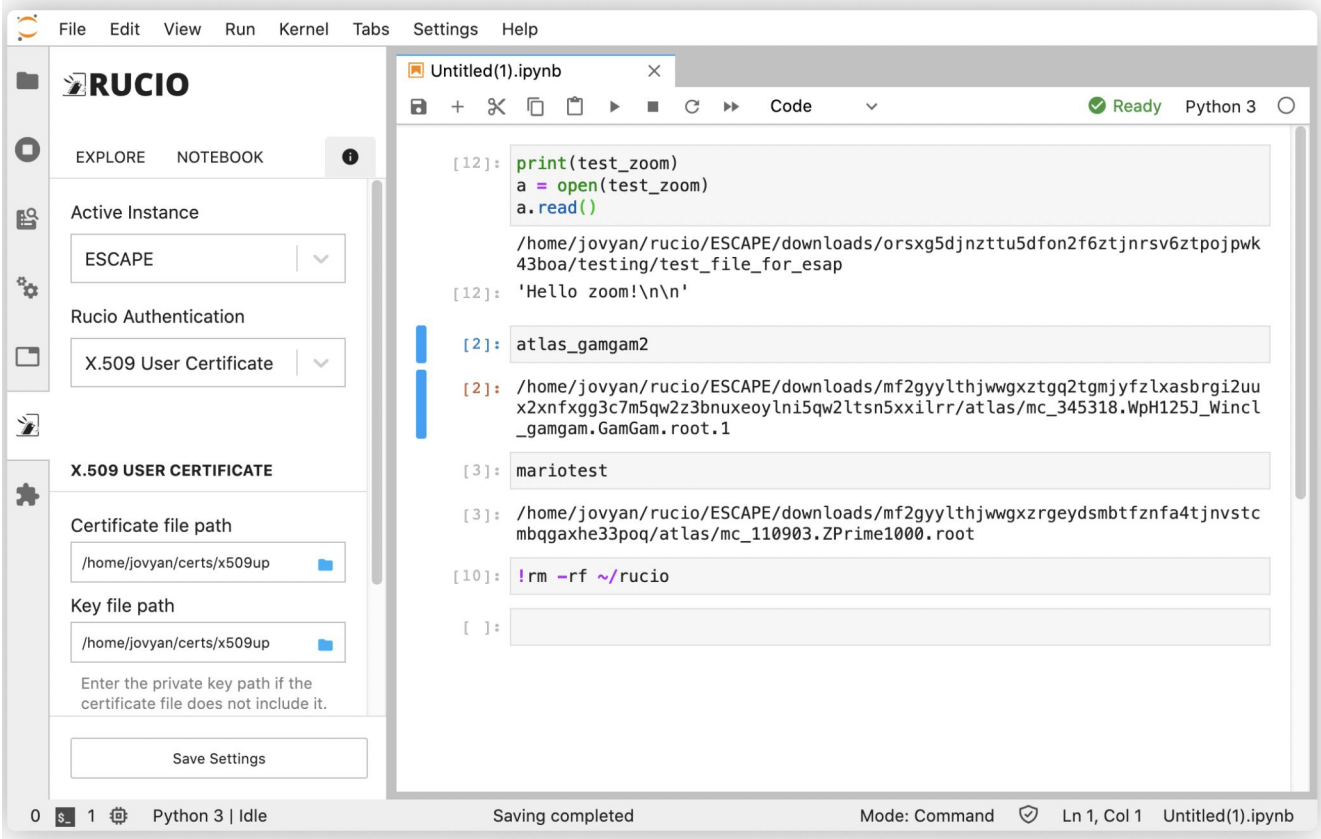
A well-known tool for all of us (data analysis and visualisation) is the Jupyter notebook.

JupyterLab is a suite of tools and features that allow interacting with multiple elements in a single view. And do the computation, of course.



The RUCIO extension for JupyterLab

The JupyterLab RUCIO extension allows to authenticate and interact with the datasets from the web UI.
Making much easier the exploration and analysis of samples in the Data Lake infrastructure.



ATLAS Open Data

We deploy the resources on the Internet.
In a nutshell, they are a series of

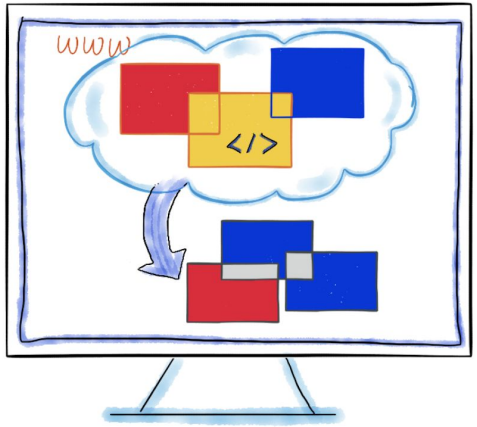
- Data samples in ROOT n-tuple format
- Software and Jupyter Notebooks in Python and C++ to analyse the samples and produce physics analysis
- JavaScript (JS) applications to produce cut-and-count analysis
- Virtual Machines with several Linux-based OS and ROOT CERN analysis framework
- GitHub & GitLab repositories
- GitBooks to document the several possible activities that can be performed

Data & Tools Repository

Here you have in a single place all the necessary pieces in order to start you physics analysis in a more complete way. Look into the data like an ATLAS particle physicist!

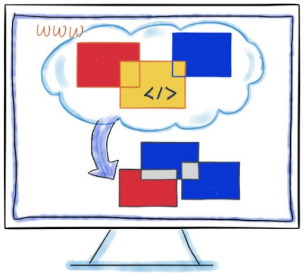
In this section, you can find where to download:

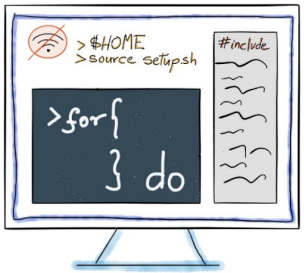
- The complete collection of available datasets
- The different analysis software
- The virtual machines to perform physics searches

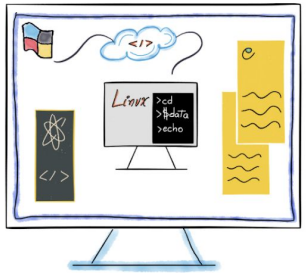


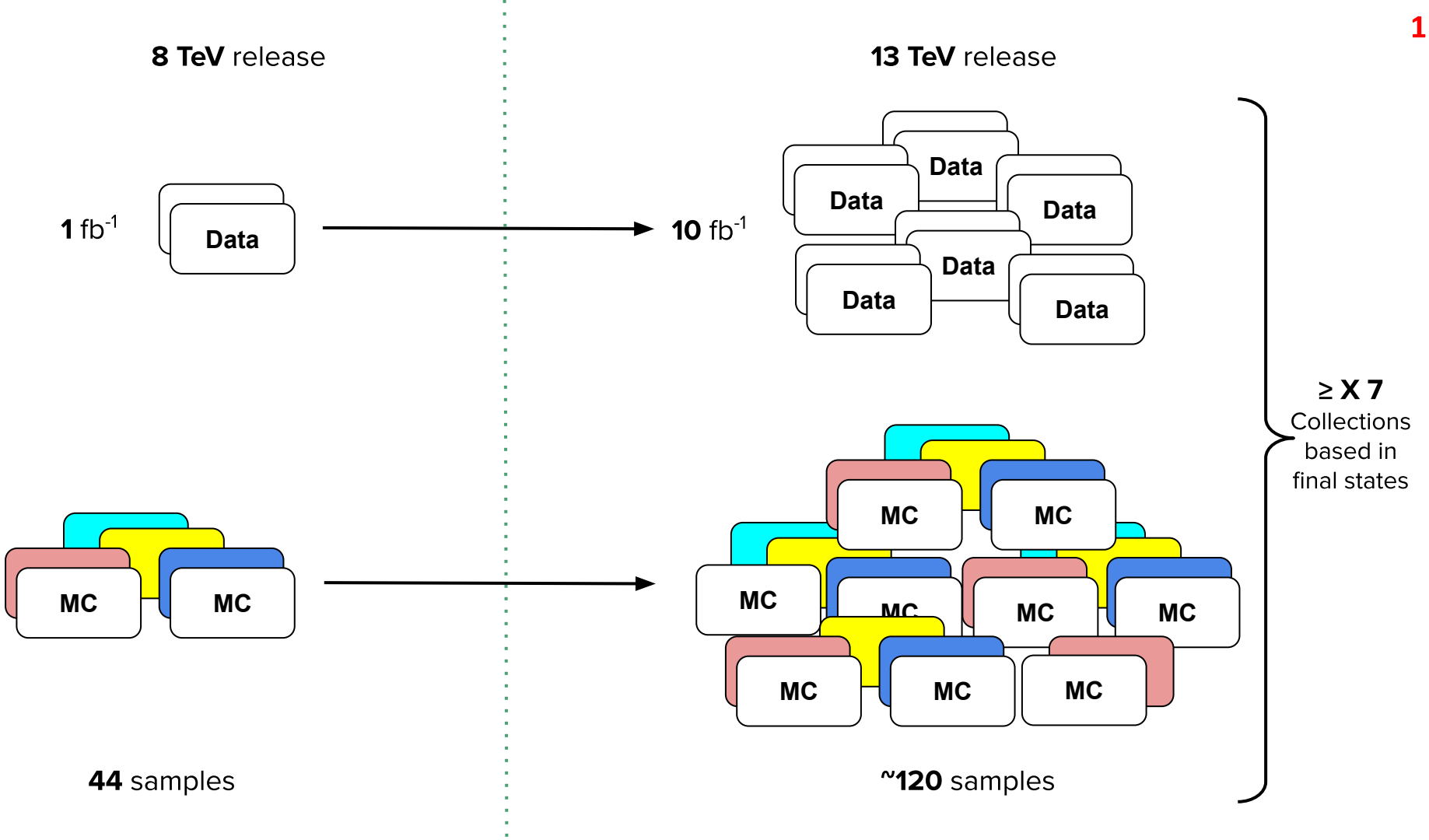
Downloads

Datasets	Software	Virtual Machines



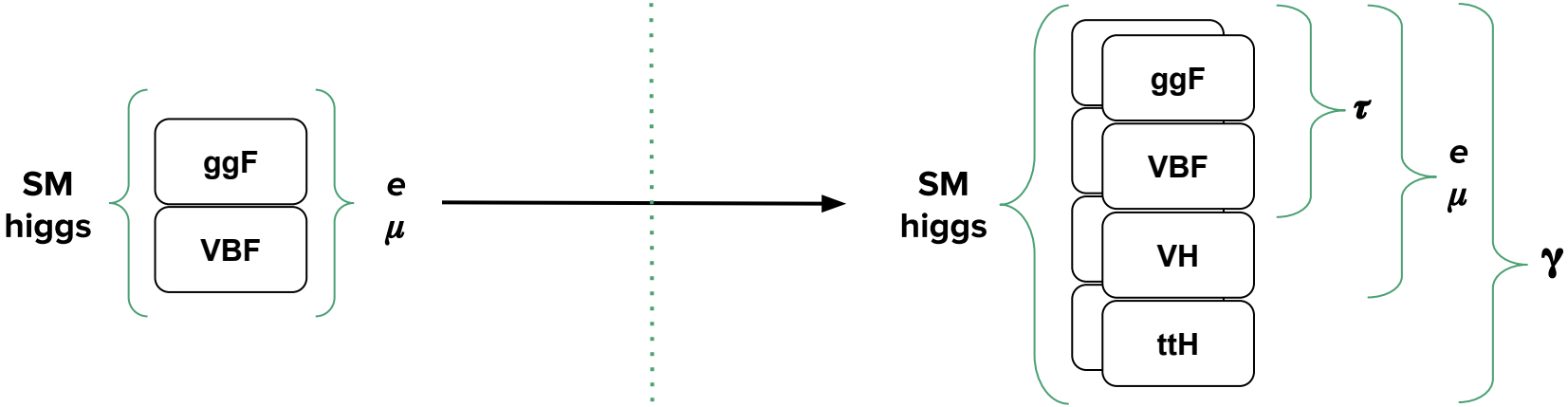


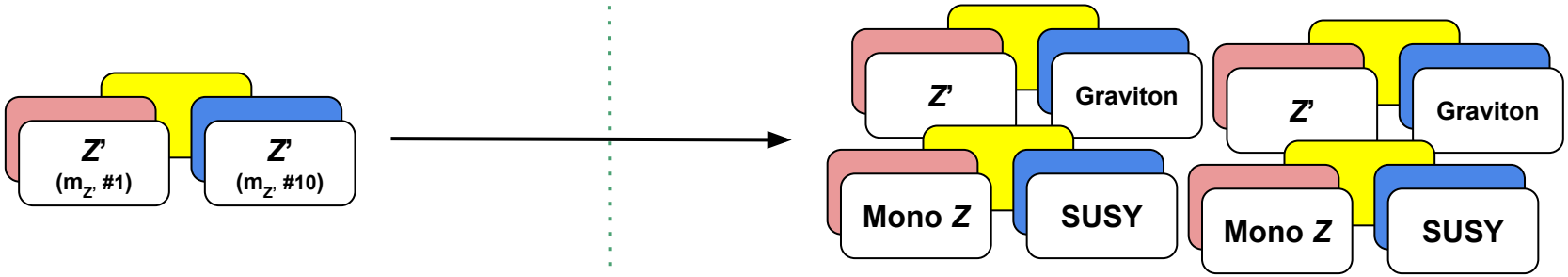



8 TeV release
13 TeV release
1 fb-1
Data
10 fb-1
Data
Data
Data
Data
Data
Data
MC
MC
MC
MC
MC
MC
MC
MC
MC
MC
≥ X 7
Collections
based in
final states
44 samples
~120 samples

8 TeV MC signals

13 TeV MC signals

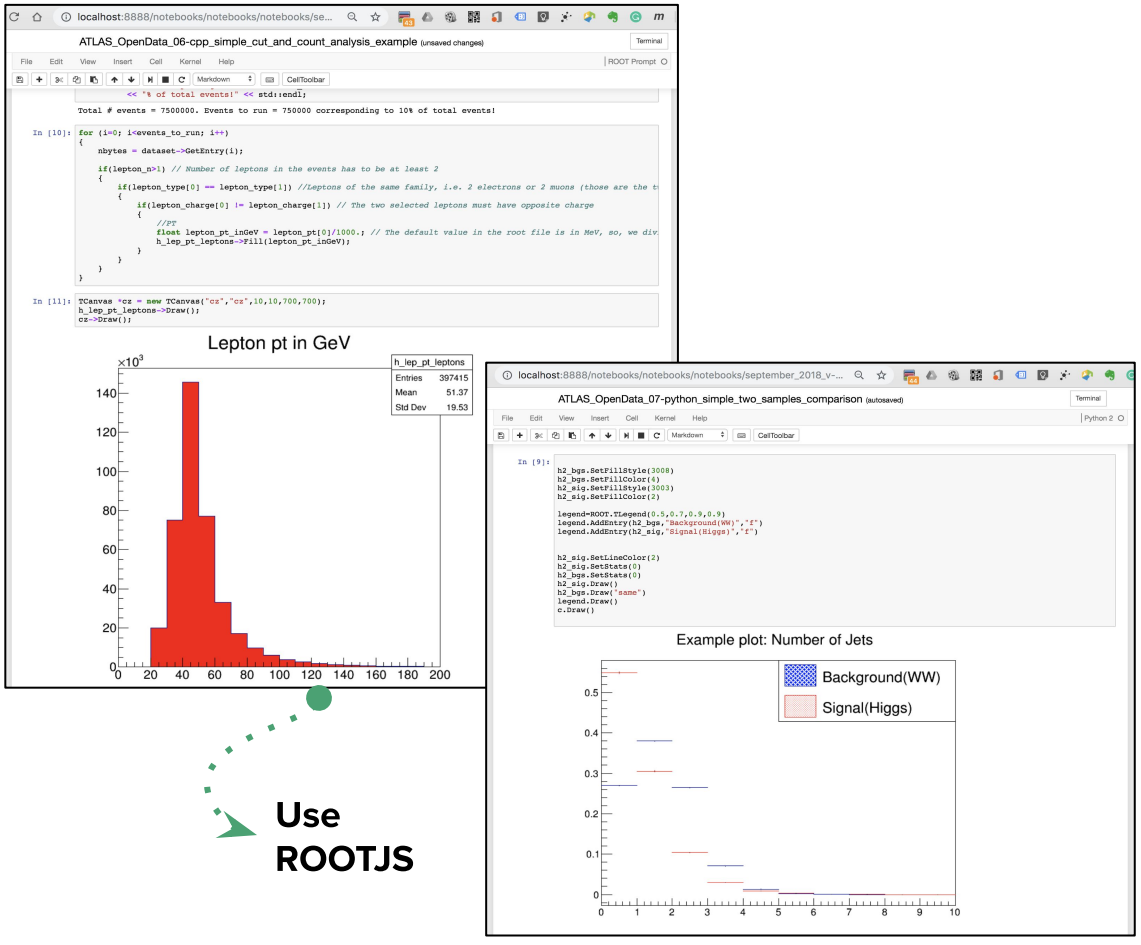




14 BSM samples

≥ 50 BSM samples

Jupyter Notebooks



Jupyter notebooks can run ROOT commands

- We produce a series of examples for basic training on the usage of the notebooks, reading of the samples and plotting simple analysis.
- The notebooks use both the Python and the C++ ROOT kernel to produce results that can be adjusted by teachers and trainers.
- The notebooks can read the samples directly from the Internet (using http protocol) or run local (if there is limited Internet access)

The pieces together:

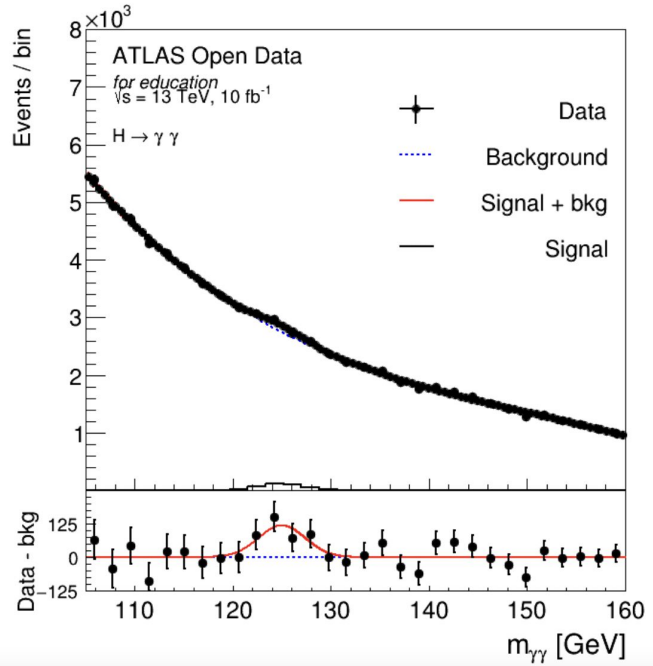
ATLAS Jupyter Notebooks and JupyterLab RUCIO extension

The ATLAS Open Data as a test field

```
In [8]: for i in range(len(hyy_0)):
            print(i, hyy_0[i])

0 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
1 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
2 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
3 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
4 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
5 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
6 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
7 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
8 /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjsxqzlsmnuxgzk7jb
```

```
In [10]: show_image('histograms/hist_mYY_bin1.png')
```



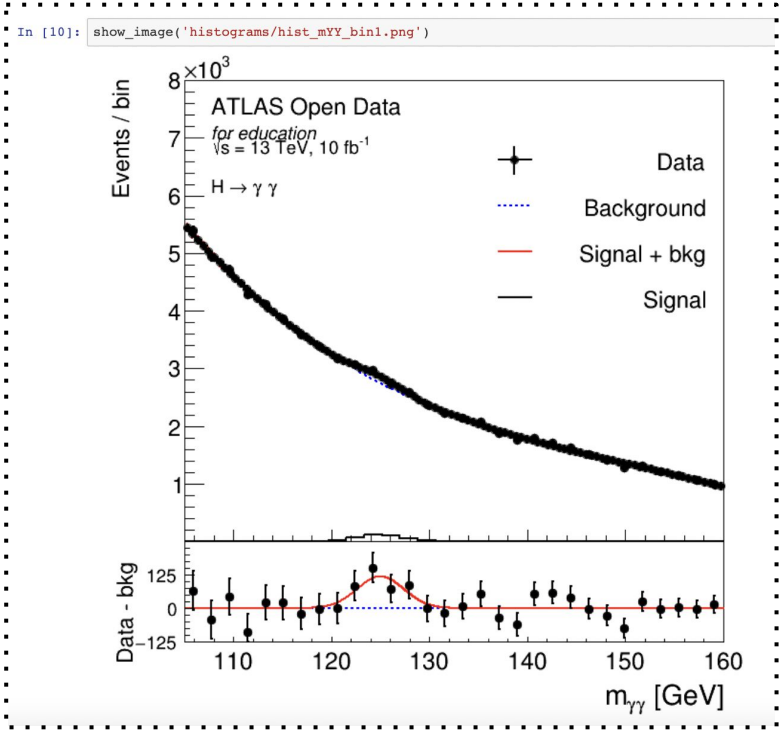
Once the Open Data datasets are registered in RUCIO (*like is the case for the current 13 TeV samples, like those in this Higgs into two photons example, thanks to Stephane)* they can be downloaded and read, using the JupyterLab extension, including search features

Example in nbviewer.jupyter.org

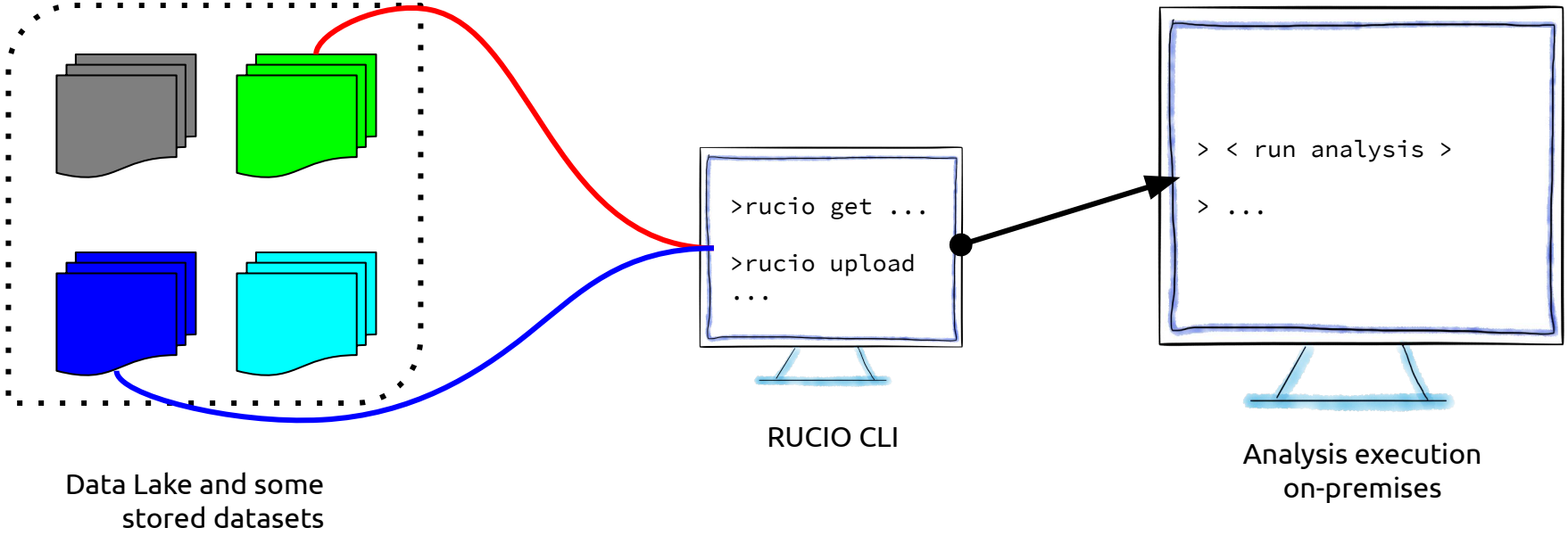
The ATLAS Open Data as a test field

Much more computational complex particle physics analysis already exist and they will be used/converted and improved when needed so to be a proper set of analysis examples as close as possible to “real” analysis, using the existing publicly available data

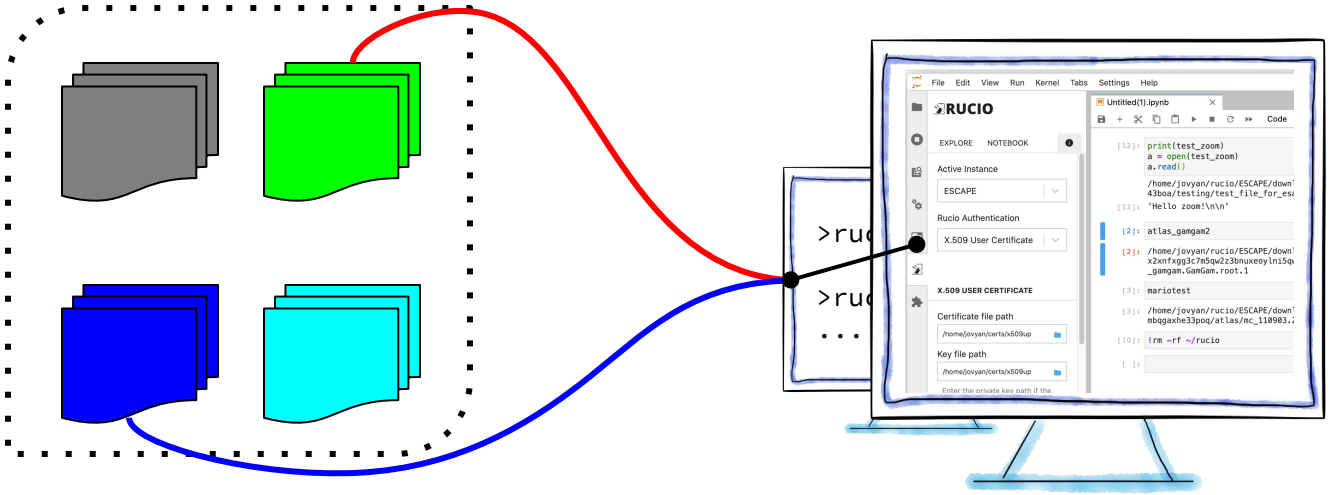
More in Opendata.atlas.cern - documentation 13 TeV - physics.



Recap



Traditional interaction with samples

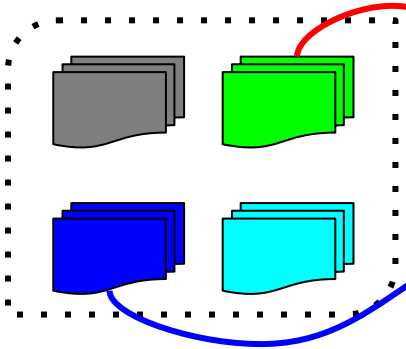


Authentication, discovery and download of the samples using the Web UI

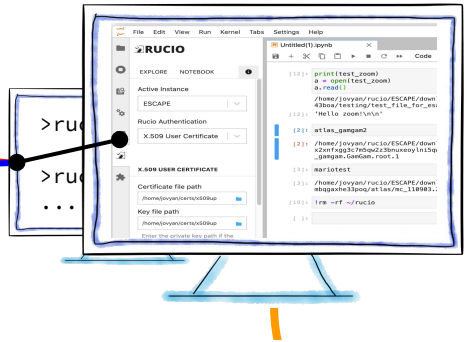
Data Lake and stored Open Data datasets

RUCIO JupyterLab extension

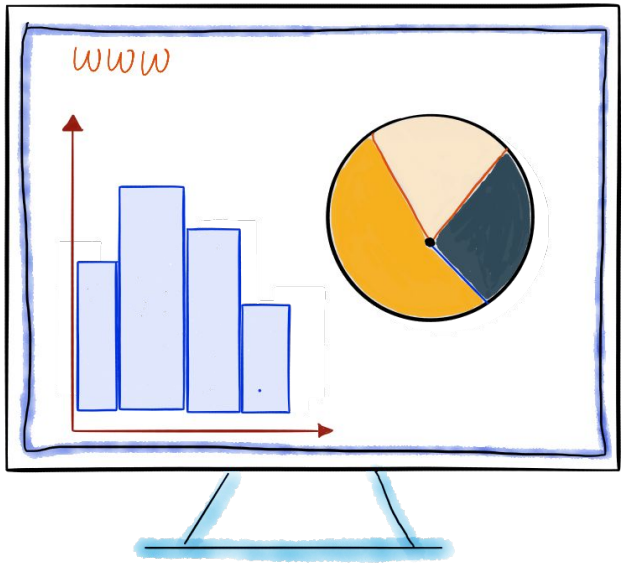
RUCIO+JupyterLab proposal for end users



Data Lake and stored
ATLAS OD datasets



RUCIO JupyterLab
extension on ATLAS OD
notebooks for testing



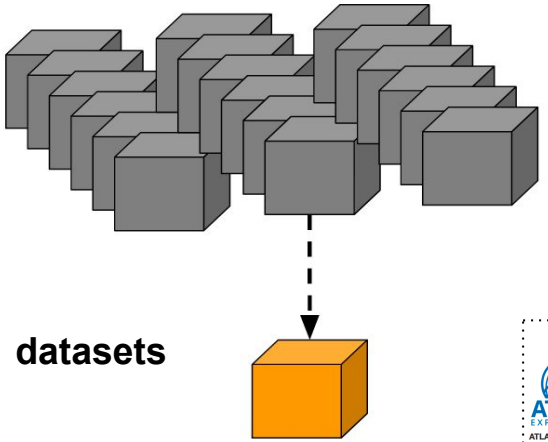
Analysis code, results and
visualisation

A view of the service

Summary

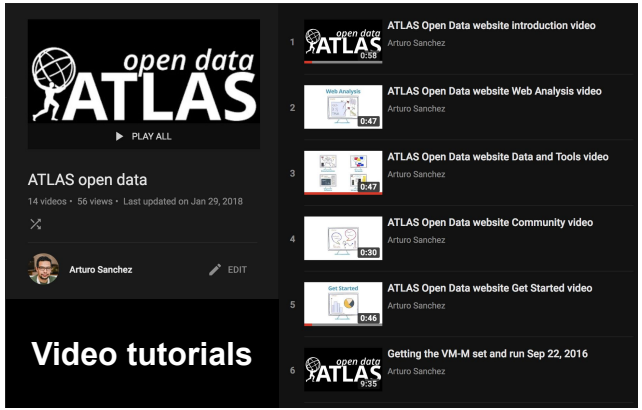
The job now is the testing, consolidation and use of the mentioned resources in a consistent way that resembles a single service + analysis of real experimental data.

Backup



datasets

The ATLAS Open Data project (more here) aims to release real and simulated Data, together with Open Source software resources to analyse those samples.
As well as documentation in several useful formats!

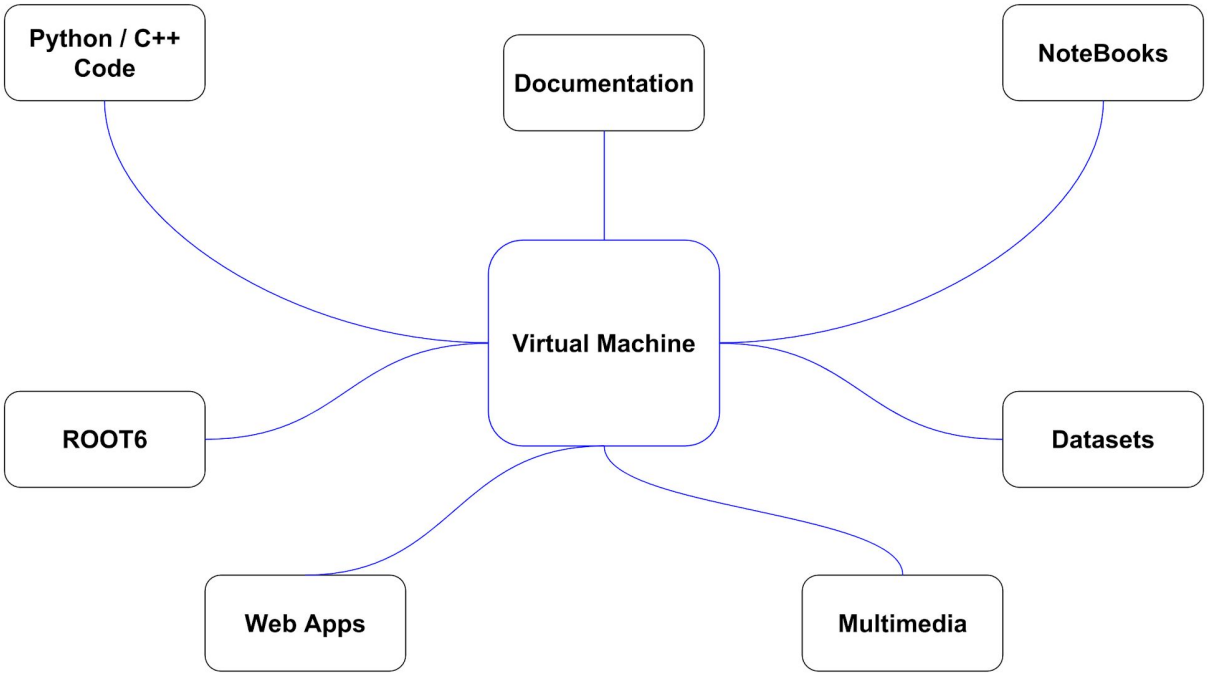


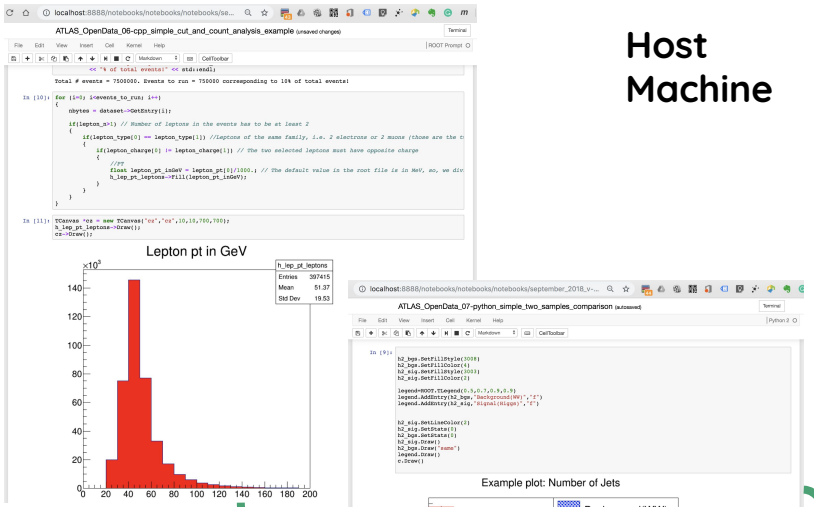
http://opendata.atlas.cern/new

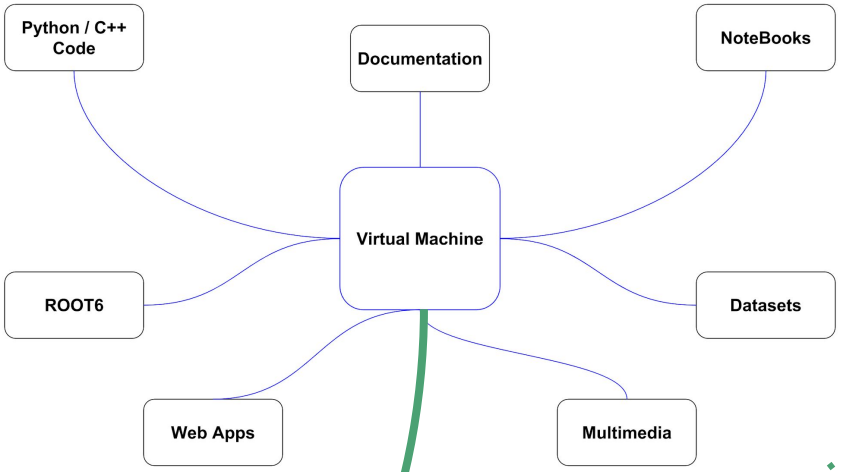
Virtual Machines as Servers

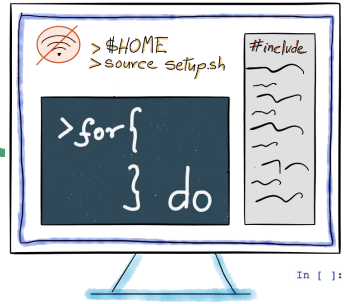
Quick view of the anatomy of the ATLAS Open Data VM

- Based in a Linux-kind OS with standard graphical UI
- The OS in enhanced with *all* ROOT's needed libraries and dependencies
- ROOT5 or ROOT6 analysis framework and IPython
- The Open **Datasets and Software** analysis frameworks
- Jupyter-notebook technology and Examples Notebooks
- Documentation in form of PDFs and Video tutorials.









Uses as a Server

Jupyter notebooks can run ROOT commands and other Python libraries and tools

- The notebooks use both the Python and the C++ ROOT kernel to produce results using the VM as a server, teaching as well the principles of Cloud and Distributed Computing.

```
In [ ]: import sys
        !{sys.executable} -m pip install --upgrade --user pip
        !{sys.executable} -m pip install -U numpy pandas uproot matplotlib keras scikit-learn --user
```

```
In [ ]: import os
        import csv
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        from pandas import read_csv
        from matplotlib import pyplot
```