CERN openlab

Data Lake as a Service

It’s a lake.. but for data.. presented as a service

Openlab Summer Students Lightning Talks 2021

Muhammad Aditya Hilmy <aditya.hilmy@cern.ch>

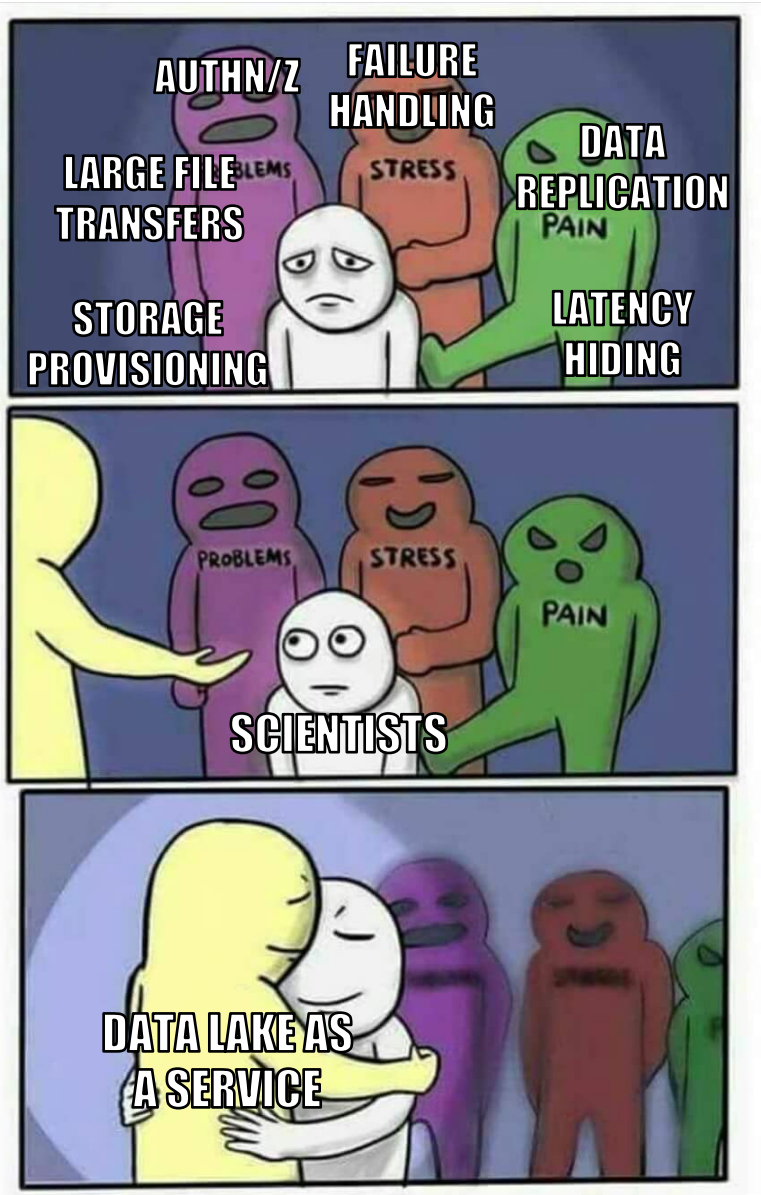
06 / 09 / 2021

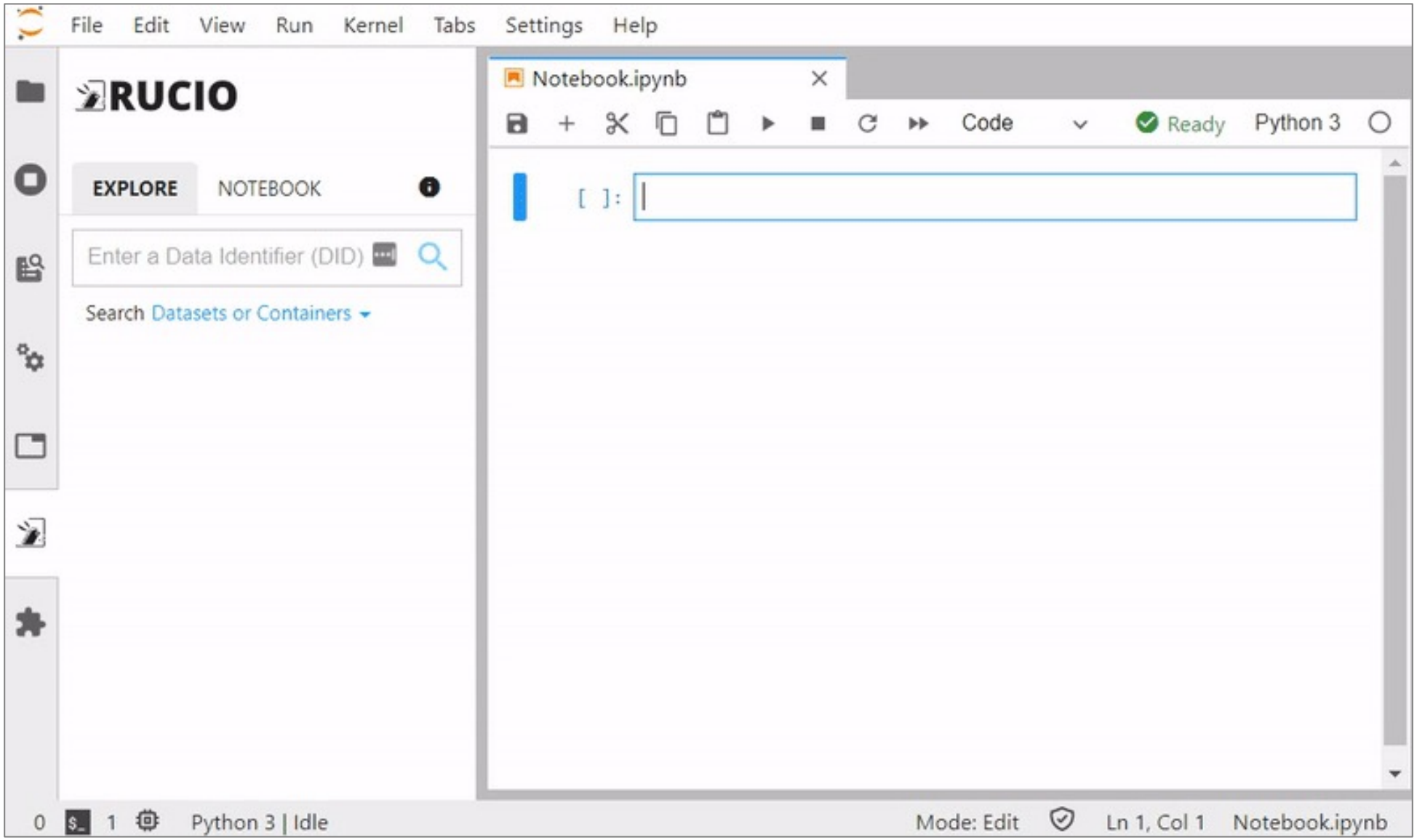
A bit of context

- We will have HL-LHC and other experiments coming online.
- Data volume expected to increase by >10x.
- We need to think about how to store and manage the data.
- The Data Lake is a place where experiments can ‘dump’ their data.
- …and scientists can ‘fish’ data from.
- The challenge: making sure the scientists can ‘fish’ easily.

Making data fishing easier

- The Data Lake has a lot of moving parts.
- The goal of the service is to hide the complexities of the Data Lake from the scientists.
- This way, scientists can focus their time on doing science-y stuff.

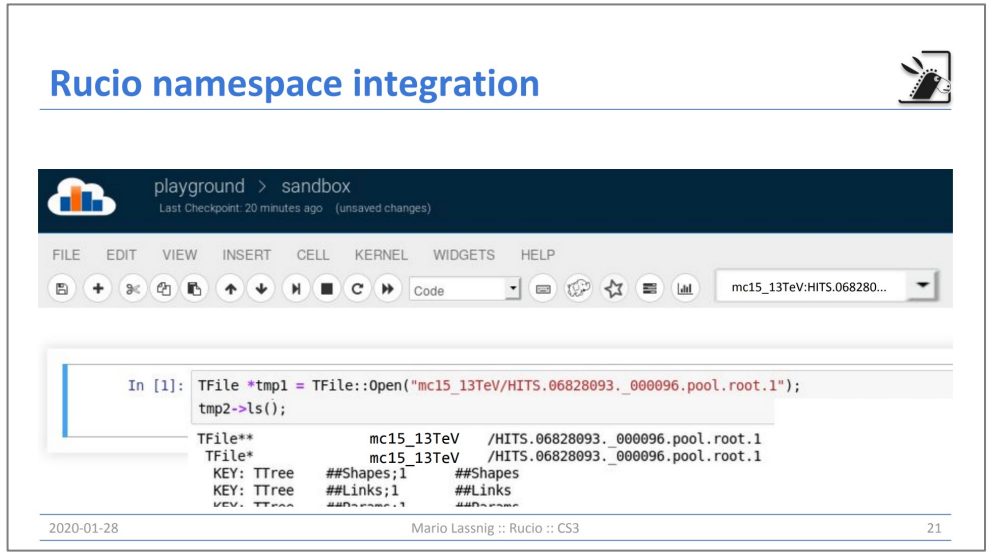


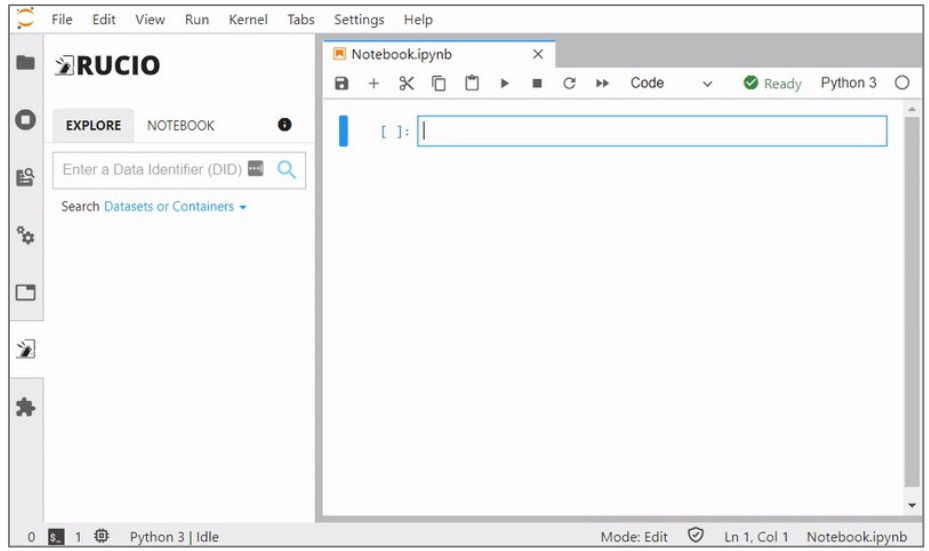


Full demo video: https://youtu.be/AxzAsXTEaxw

A humble beginning

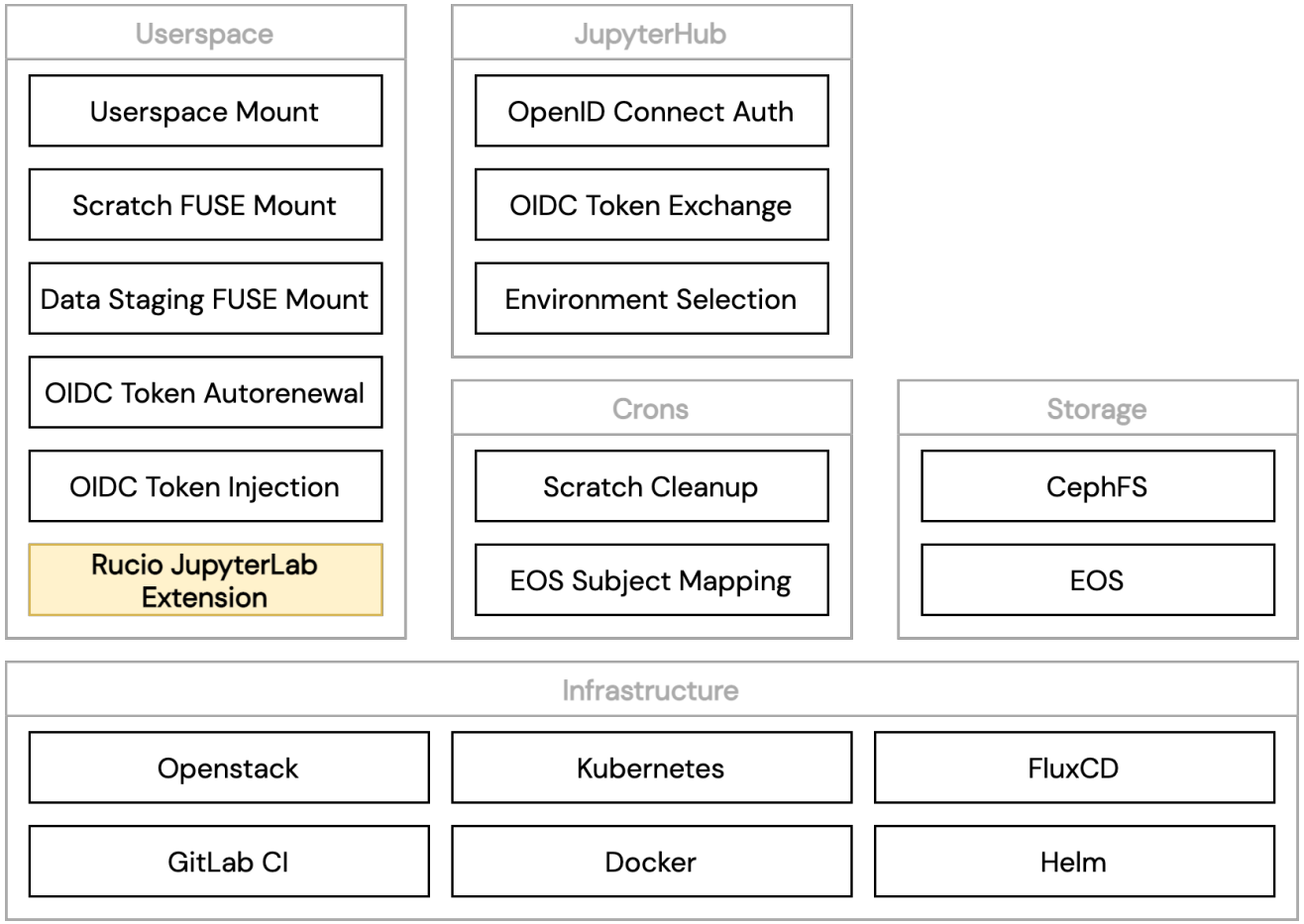
- Started as an idea presented on CS3 2020 by the Rucio team [1]
- Developed “Rucio JupyterLab Extension” as a part of Google Summer of Code 2020 [2,3]
- Deployed the extension as “Data Lake as a Service” as a part of Openlab Online (☹) Summer Students Programme 2021





[1] https://indico.cern.ch/event/854707/contributions/3680520/
[2] https://hepsoftwarefoundation.org/gsoc/2020/proposal_SWAN_RUCIO_integration.html
[3] https://summerofcode.withgoogle.com/archive/2020/projects/6320550214893568/

The DataLake as a Service



A bit of early mishaps

I broke production system on my first week :)

25-06-2021 11:31:22 - Spyridon Trigazis Additional comments (Customer View)

Hello,

It was Muhammad Aditya Hilmy the deleted both pools.
2021-06-22 11:58:40.630 13 INFO neutron.wsgi [req-ed93c560-efe3-4070-a94f-e0366f218844 muhilmy 586f1e21-4512-4480-890b-756cf4b7facb - default default] 137.138.121.172,10.100.1.0 "DELETE /v2.0/lbaas/pools/bf36789b-a9b5-4dc9-aec8-c3ca41e13abf HTTP/1.1" status: 204 len: 149 time: 1.0106199

2021-06-22 11:58:11.909 13 INFO neutron.wsgi [req-df990818-9eb1-4e21-916a-cd028e229016 muhilmy 586f1e21-4512-4480-890b-756cf4b7facb - default default] 137.138.121.172,10.100.1.0 "DELETE /v2.0/lbaas/pools/58603cd6-cd6d-468b-84a5-333490415ee0 HTTP/1.1" status: 204 len: 149 time: 0.9526761

I will check with my colleagues if we can manually recreate the pools, but heat operations will fail since the pool IDs witll be new.

Cheers,
Spyros

Use Cases

- Data discovery and access
- Data analysis
  - Use the files directly in the JupyterLab interface to create plots.
- Data preparation and processing
  - Use the service to preprocess data in the Data Lake, and once done, upload it back to Rucio.
- Data preservation
  - Use the service to produce data and reupload them to the Data Lake
- Submitting jobs to external service (remote computing)
  - Users can use the convenience of the extension to browse data in Rucio and access the file PFN directly from the notebook code.

Desktop Data Lake-as-a-Service

Location-aware Remote Configuration

More Kernel Support

Remote Configuration from Notebook File

Data Lake as a Service

Token Support for Direct Download

File Upload Functionality

Integration with SWAN

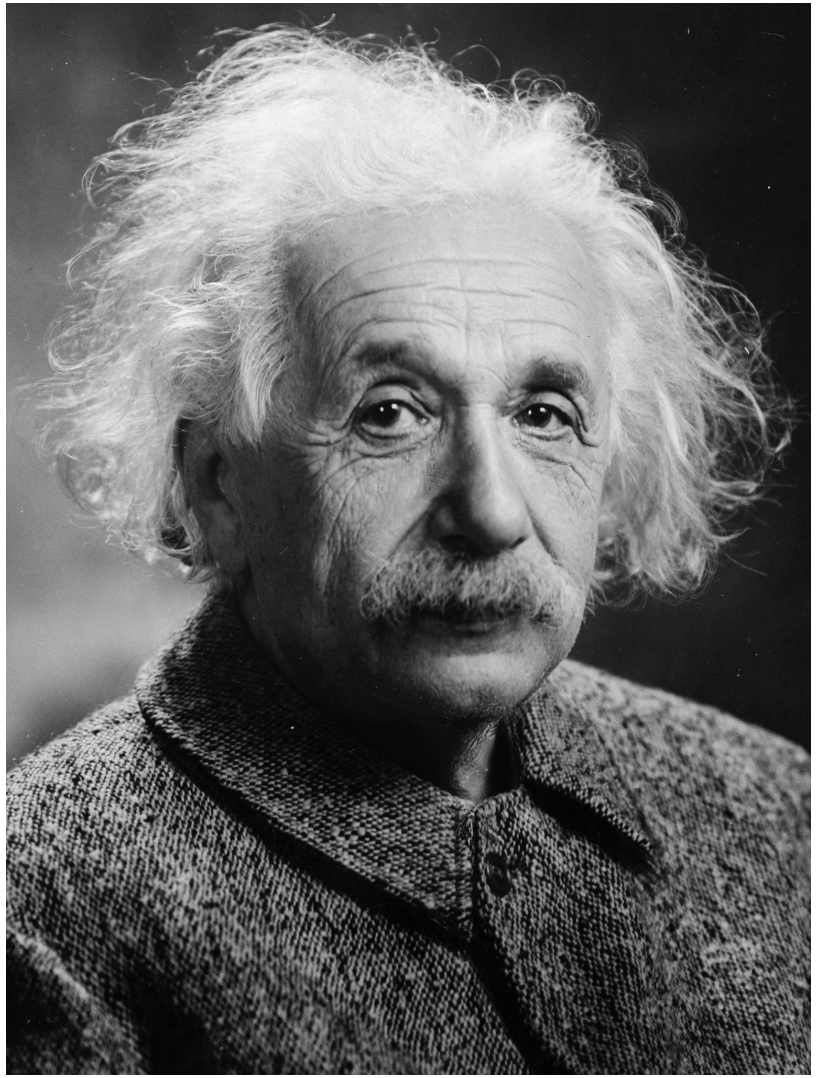
Integration with ESCAPE ESAP

Latency Hiding Layer

Acknowledgements

Here are the people who've helped me throughout the journey:

Riccardo Di Maria (IT-SC-RD)
Xavier Espinal (IT-SC-RD)
Alba Vendrell Moya (IT-SC-RD)
Rizart Dona (IT-SC-RD)
David Smith (IT-SC-RD)
Cristian Contescu (IT-ST-PDS)
Elvin Alin Sindrilaru (IT-ST-PDS)
Andreas Joachim Peters (IT-ST-PDS)
Mihai Patrascoiu (IT-ST-PDS)
Enrico Bocchi (IT-ST-GSS)
Diogo Castro (IT-ST-GSS)
Mario Lassnig (EP-ADP-CO)
Martin Barisits (EP-ADP-CO)



“Thank you”

- Albert Einstein, at some point in time

Attributions:
The SpongeBob GIF is a copyright of Viacom International
The stick figure is taken from XKCD (Randall Munroe)
This picture of Albert Einstein is taken from Wikimedia Commons

QUESTIONS?

mhilmy@hey.com

Muhammad Aditya Hilmy didithilmy