Jupyter Notebooks as scientific gateways to access cloud computing and distributed storage

Fernando Aguilar
IFCA (CSIC-UC)
aguilarf@ifca.unican.es

XDC

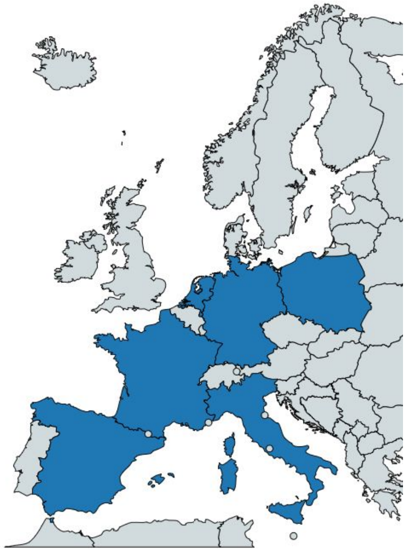
- The eXtreme DataCloud is a software development and integration project
- Develops scalable technologies for federating storage resources and managing data in highly distributed computing environments
    - Focus efficient, policy driven and Quality of Service based DM
- The targeted platforms are the current and next generation e-Infrastructures deployed in Europe
    - European Open Science Cloud (EOSC)
    - The e-infrastructures used by the represented communities
- Addresses the EINFRA-21-2017 (b)-2: “Computing e-infrastructure with extreme large datasets”
    - Deal with heterogeneous datasets
    - Bring to TRL8 and include in a unified service catalogue services and prototype at least at TRL6

XDC Consortium

eXtreme DataCloud

ID	Partner	Country	Represented Community	Tools and system
1	INFN (Lead)	IT	HEP/WLCG	INDIGO-Orchestrator
2	DESY	DE	Research with Photons (XFEL)	dCache
3	CERN	CH	HEP/WLCG	EOS, DYNAFED, FTS, RUCIO
4	AGH	PL		ONEDATA
5	ECRIN	[ERIC]	Medical data	
6	UC	ES	Lifewatch	
7	CNRS	FR	Astro [CTA and LSST]	
8	EGI.eu	NL	EGI communities	

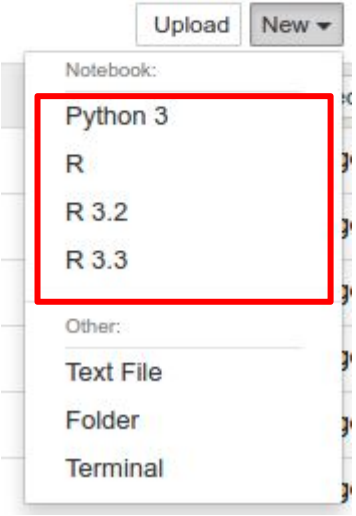
- 8 partners, 7 countries
- 6 research communities represented + EGI
- XDC Total Budget: 3.07Meuros

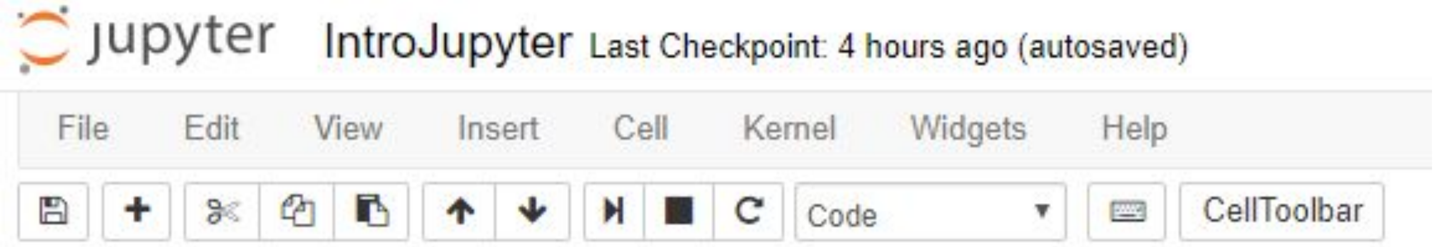


What is Jupyter?

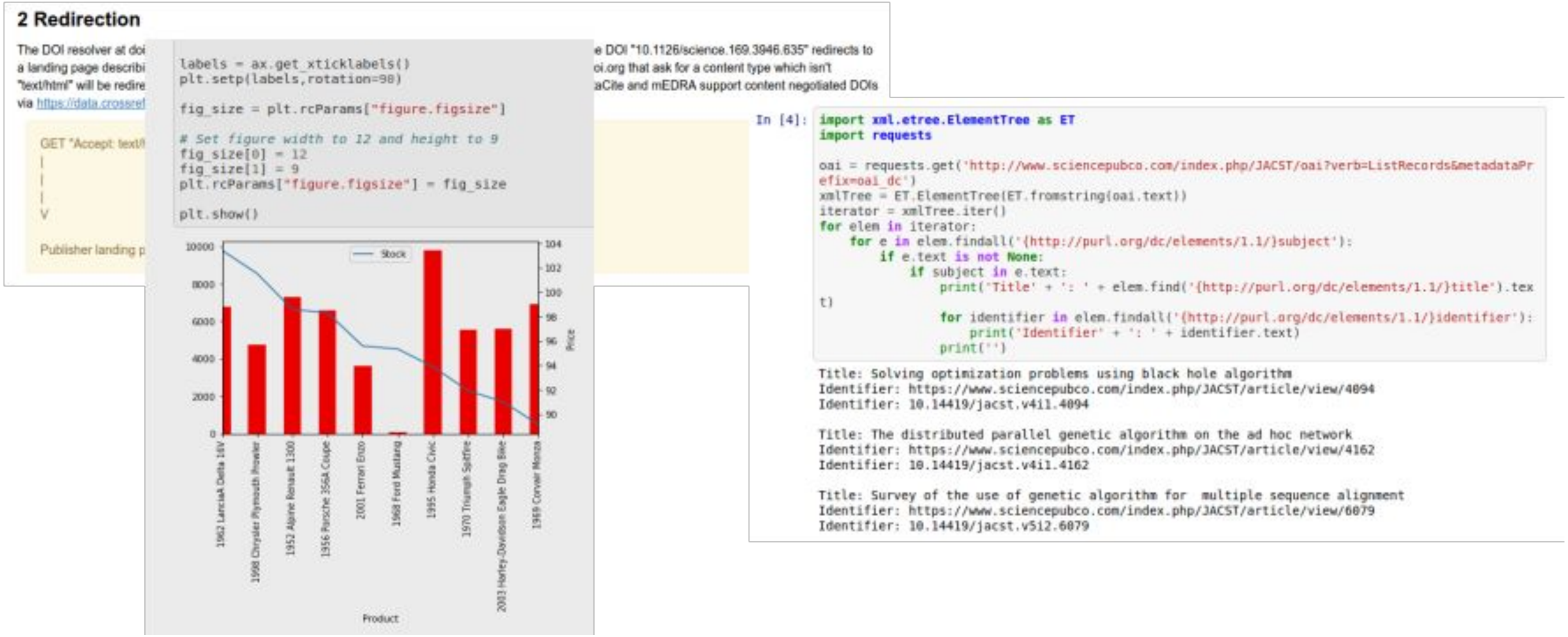
- Notebooks are documents produced by Jupyter Notebook App.
- Contains both source code and enriched text.
- Understood by humans, actionable by machines: scripts, data analytics, etc.
- Useful for teaching, user applications.
- Different kernels/programming languages.

Juptyer: Main Menu





Applications



2 Redirection

The DOI resolver at doi … e DOI "10.1126/science.169.3946.635" redirects to a landing page describi … oi.org that ask for a content type which isn't "text/html" will be redire … aCite and mEDRA support content negotiated DOIs via https://data.crossref

```
GET "Accept: text/
|
|
|
V

Publisher landing p
```

```
labels = ax.get_xticklabels()
plt.setp(labels,rotation=90)

fig_size = plt.rcParams["figure.figsize"]

# Set figure width to 12 and height to 9
fig_size[0] = 12
fig_size[1] = 9
plt.rcParams["figure.figsize"] = fig_size

plt.show()
```

```
In [4]: import xml.etree.ElementTree as ET
        import requests

        oai = requests.get('http://www.sciencepubco.com/index.php/JACST/oai?verb=ListRecords&metadataPr
        efix=oai_dc')
        xmlTree = ET.ElementTree(ET.fromstring(oai.text))
        iterator = xmlTree.iter()
        for elem in iterator:
            for e in elem.findall('{http://purl.org/dc/elements/1.1/}subject'):
                if e.text is not None:
                    if subject in e.text:
                        print('Title' + ': ' + elem.find('{http://purl.org/dc/elements/1.1/}title').tex
        t)
                        for identifier in elem.findall('{http://purl.org/dc/elements/1.1/}identifier'):
                            print('Identifier' + ': ' + identifier.text)
                        print('')
```

```
Title: Solving optimization problems using black hole algorithm
Identifier: https://www.sciencepubco.com/index.php/JACST/article/view/4094
Identifier: 10.14419/jacst.v4i1.4094

Title: The distributed parallel genetic algorithm on the ad hoc network
Identifier: https://www.sciencepubco.com/index.php/JACST/article/view/4162
Identifier: 10.14419/jacst.v4i1.4162

Title: Survey of the use of genetic algorithm for  multiple sequence alignment
Identifier: https://www.sciencepubco.com/index.php/JACST/article/view/6079
Identifier: 10.14419/jacst.v5i2.6079
```

Jupyter Hub

- JupyterHub brings the power of notebooks to groups of users.
- It gives users access to computational environments and resources without burdening the users with installation and maintenance tasks.
- Users with their own workspace.
- Features:
    - **Customizable** - JupyterHub can be used to serve a variety of environments. It supports dozens of kernels with the Jupyter server.
    - **Flexible** - Authentication is pluggable, supporting a number of authentication protocols (such as OAuth and GitHub).
    - **Scalable** - JupyterHub is container-friendly.
    - **Portable** - JupyterHub is entirely open-source.

Docker + AAI

- JupyterHub adopts DockerSpawner to deploy the user workspace.
- A Docker image can be configured with any software/package required by the user.
- It can be configured to perform any kind of actions when the user logs in, logs out, etc.
- The Authentication and Authorization is compatible with multiple social IDs (Github, Google, etc.).
- It can also be configured with AAI standards, like OpenID Connect.
- Information about the user (username, tokens) can be sent to the docker container deployed as workspace for the user.

OpenID-Connect - INDIGO IAM (Identity and Access Management)

Flexible authentication support (SAML, X.509, OpenID Connect, username/ password, …)

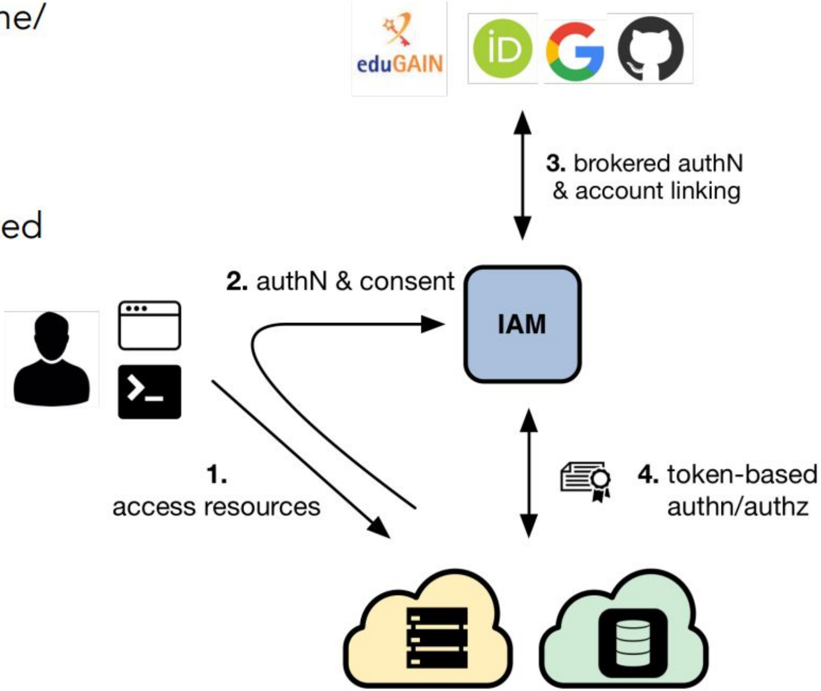
Account linking

Registration service for moderated and automatic user enrollment

AUP enforcement support

Mobile-friendly organization management tools

Easy integration in off-the-shelf components thanks to OpenID Connect/OAuth


eduGAIN
iD
IAM
1. access resources
2. authN & consent
3. brokered authN & account linking
4. token-based authn/authz

Slide: Andrea Ceccanti

Scientific Gateway

- Integrating different components using the same AAI, a new complete environment can be deployed, including all the required components for researchers.
- Data gathering, configuration, programming, visualization: JupyterHub
- Computing needs: PaaS Orchestrator (jobs submission)
- Storage needs: Onedata. Cloud storage.
- INDIGO IAM and OIDC standard are the “glue” to integrate the different elements.

INDIGO - PaaS Orchestrator

Kind of “batch system” to send “jobs” to Cloud Computing resources.

Collects high-level deployment requests and translate them into action to coordinate resources interacting with the underlying cloud infrastructures.

Allows to implement workflows with different steps (data ingestion, data processing, etc.).

New features are developed in XDC project.

Onedata

Distributed storage space to store not only data, but also customized metadata.

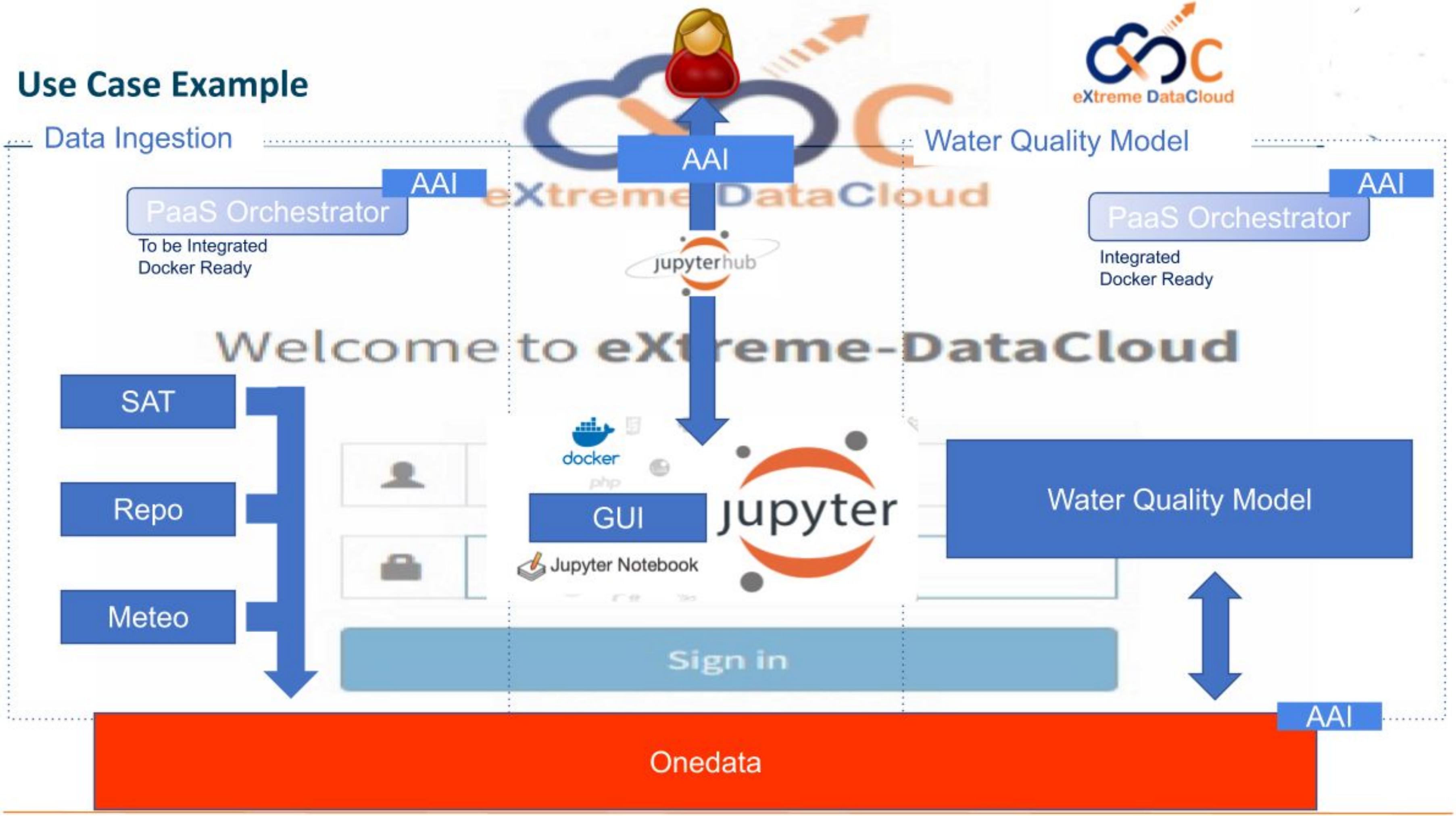
Organized in Spaces (user, can be shared), providers and zones.

One “OneZone” federates multiple providers, that can be geographically distributed.

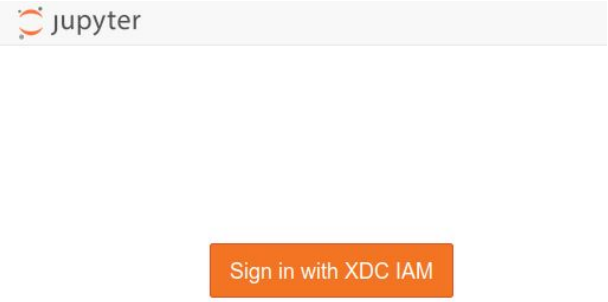
Data can be access via web, but is POSIX-compliant (directly mounted).

New features are developed in XDC project.

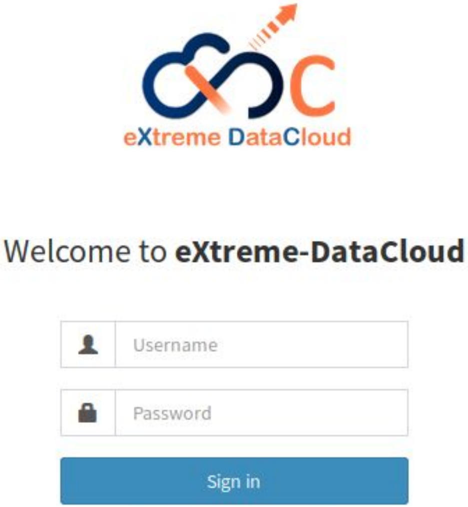
Use Case Example



Example



Sign in with XDC IAM

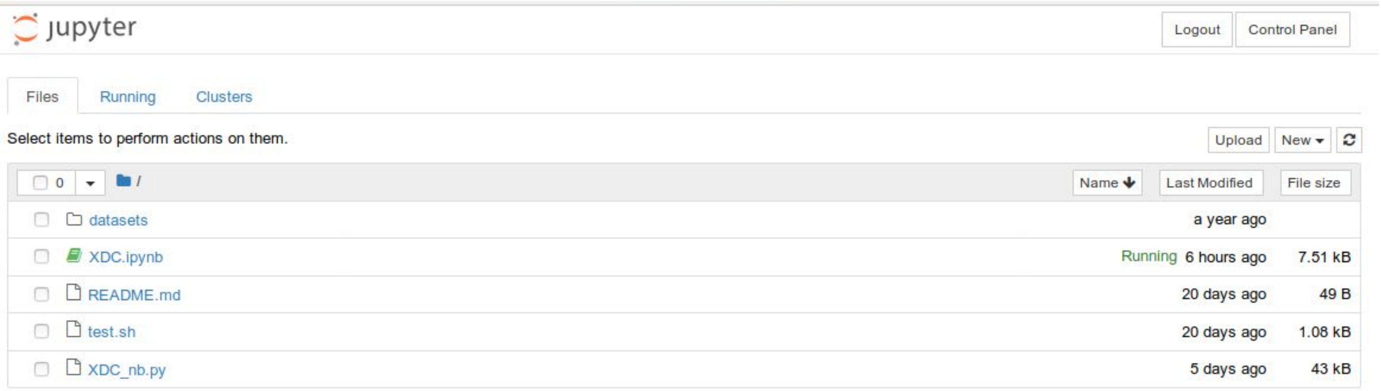


Welcome to **eXtreme-DataCloud**

Username

Password

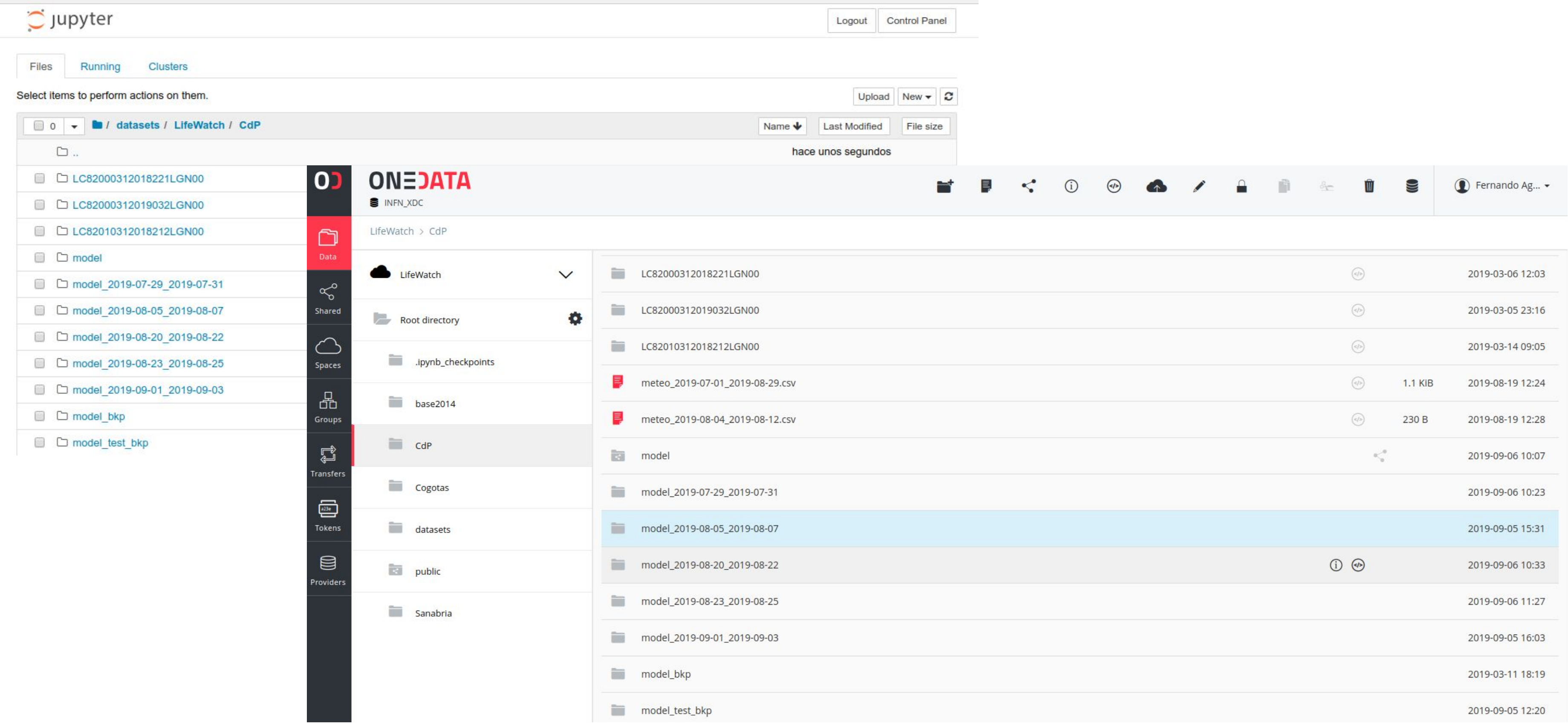
Sign in



Select items to perform actions on them.

□ 0 ▾ /	Name ↓	Last Modified	File size
□ datasets		a year ago	
□ XDC.ipynb	Running	6 hours ago	7.51 kB
□ README.md		20 days ago	49 B
□ test.sh		20 days ago	1.08 kB
□ XDC_nb.py		5 days ago	43 kB

Example



jupyter

Logout | Control Panel

Files | Running | Clusters

Select items to perform actions on them.

Upload | New ▾ | ⟳

□ 0 ▾ / datasets / LifeWatch / CdP

Name ↓ | Last Modified | File size

	Name	Last Modified
	..	hace unos segundos
□	LC82000312018221LGN00	
□	LC82000312019032LGN00	
□	LC82010312018212LGN00	
□	model	
□	model_2019-07-29_2019-07-31	
□	model_2019-08-05_2019-08-07	
□	model_2019-08-20_2019-08-22	
□	model_2019-08-23_2019-08-25	
□	model_2019-09-01_2019-09-03	
□	model_bkp	
□	model_test_bkp	

ONEDATA

INFN_XDC

Fernando Ag... ▾

LifeWatch > CdP

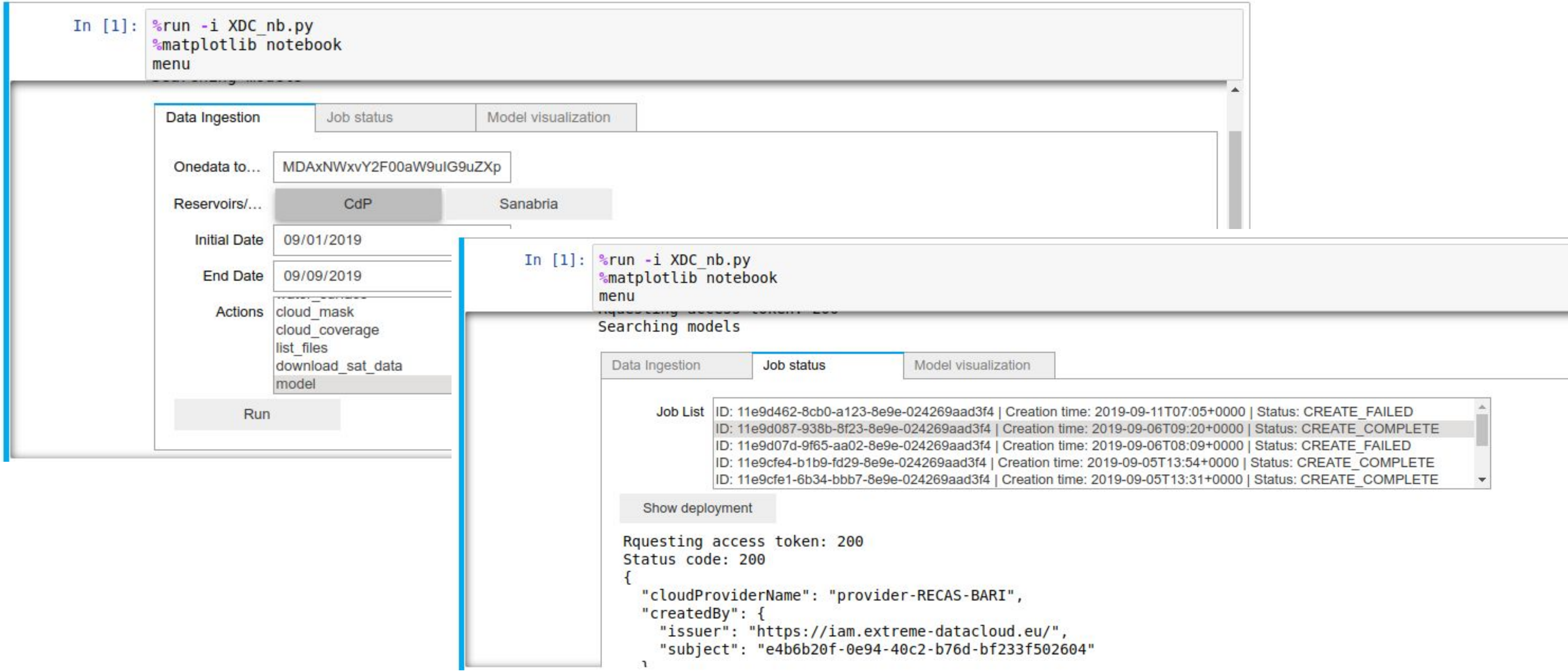
LifeWatch

Root directory

- .ipynb_checkpoints
- base2014
- CdP
- Cogotas
- datasets
- public
- Sanabria

Name	Size	Modified
LC82000312018221LGN00		2019-03-06 12:03
LC82000312019032LGN00		2019-03-05 23:16
LC82010312018212LGN00		2019-03-14 09:05
meteo_2019-07-01_2019-08-29.csv	1.1 KiB	2019-08-19 12:24
meteo_2019-08-04_2019-08-12.csv	230 B	2019-08-19 12:28
model		2019-09-06 10:07
model_2019-07-29_2019-07-31		2019-09-06 10:23
model_2019-08-05_2019-08-07		2019-09-05 15:31
model_2019-08-20_2019-08-22		2019-09-06 10:33
model_2019-08-23_2019-08-25		2019-09-06 11:27
model_2019-09-01_2019-09-03		2019-09-05 16:03
model_bkp		2019-03-11 18:19
model_test_bkp		2019-09-05 12:20

Example



```
In [1]: %run -i XDC_nb.py
        %matplotlib notebook
        menu
```

Onedata to...: MDAxNWxvY2F00aW9uIG9uZXp

Reservoirs/...: CdP | Sanabria

Initial Date: 09/01/2019

End Date: 09/09/2019

Actions: cloud_mask, cloud_coverage, list_files, download_sat_data, model

```
In [1]: %run -i XDC_nb.py
        %matplotlib notebook
        menu
```

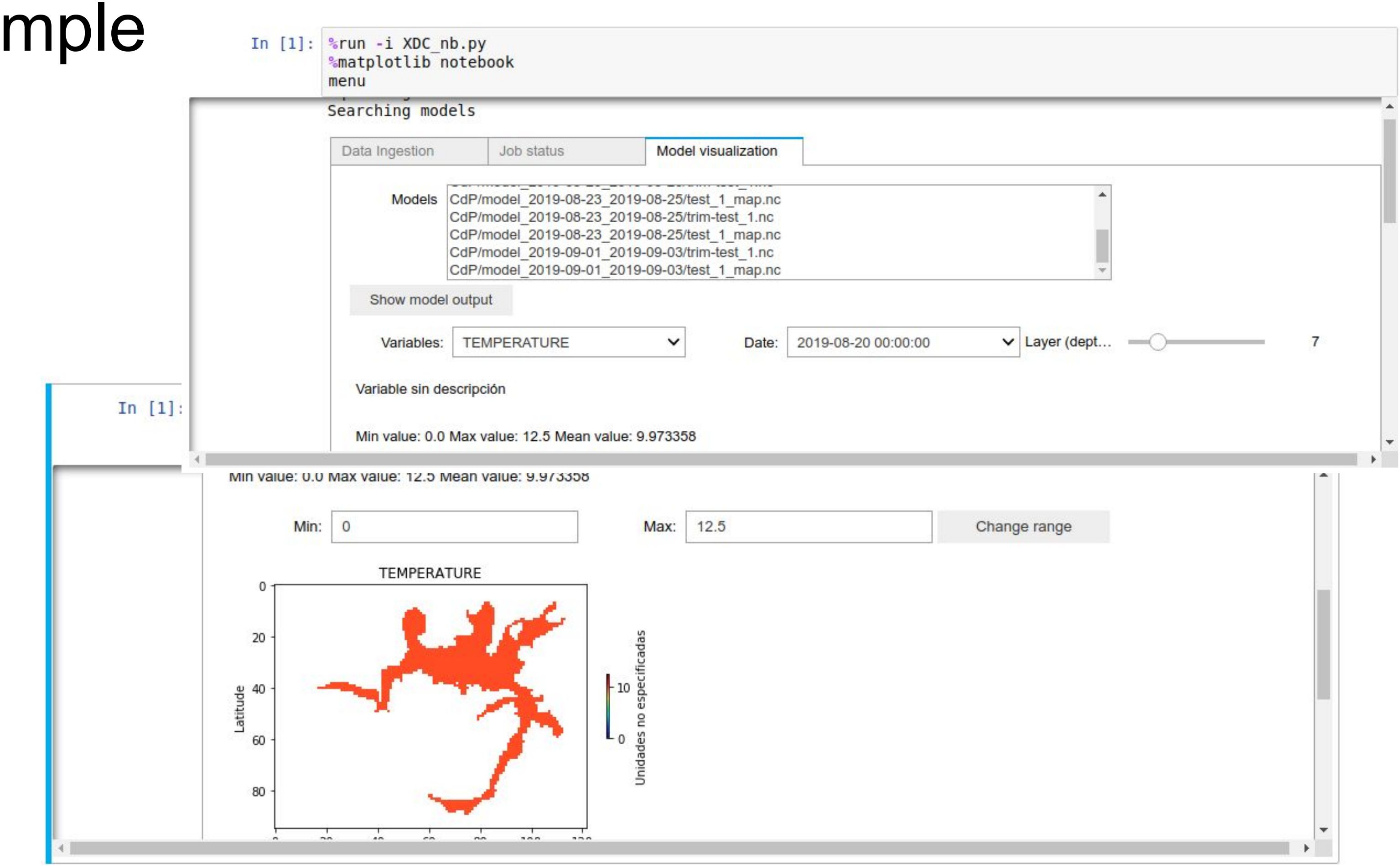
```
Searching models
```

Job List:

```
ID: 11e9d462-8cb0-a123-8e9e-024269aad3f4 | Creation time: 2019-09-11T07:05+0000 | Status: CREATE_FAILED
ID: 11e9d087-938b-8f23-8e9e-024269aad3f4 | Creation time: 2019-09-06T09:20+0000 | Status: CREATE_COMPLETE
ID: 11e9d07d-9f65-aa02-8e9e-024269aad3f4 | Creation time: 2019-09-06T08:09+0000 | Status: CREATE_FAILED
ID: 11e9cfe4-b1b9-fd29-8e9e-024269aad3f4 | Creation time: 2019-09-05T13:54+0000 | Status: CREATE_COMPLETE
ID: 11e9cfe1-6b34-bbb7-8e9e-024269aad3f4 | Creation time: 2019-09-05T13:31+0000 | Status: CREATE_COMPLETE
```

```
Rquesting access token: 200
Status code: 200
{
  "cloudProviderName": "provider-RECAS-BARI",
  "createdBy": {
    "issuer": "https://iam.extreme-datacloud.eu/",
    "subject": "e4b6b20f-0e94-40c2-b76d-bf233f502604"
```

Example



In [1]:

```
%run -i XDC_nb.py
%matplotlib notebook
menu
```

```
Searching models
```

Models
CdP/model_2019-08-23_2019-08-25/test_1_map.nc
CdP/model_2019-08-23_2019-08-25/trim-test_1.nc
CdP/model_2019-08-23_2019-08-25/test_1_map.nc
CdP/model_2019-09-01_2019-09-03/trim-test_1.nc
CdP/model_2019-09-01_2019-09-03/test_1_map.nc

Variables: TEMPERATURE

Date: 2019-08-20 00:00:00

Layer (dept… 7

Variable sin descripción

Min value: 0.0 Max value: 12.5 Mean value: 9.973358

In [1]:

Min value: 0.0 Max value: 12.5 Mean value: 9.973358

Min: 0

Max: 12.5

Thanks!

Fernando Aguilar

aguilarf@ifca.unican.es