SWAN: service for web based analysis

D. Castro, E. Tejedor, D. Piparo, P. Mato
E. Bocchi, J. Moscicki, M. Lamanna

https://swan.web.cern.ch

Dec 6th, 2017

Workshop on data analysis in large-scale research: comparing experiences in physics and biology

Introduction

Introduction

- There are analysis tools developed in CERN, but they require installation and configuration
- Some resources are only available from within CERN network
  - And remote connection might not be ideal
- External services, like IBM Bluemix or Mybinder, provide analysis services but lack some advanced features
  - Like software packages and integration with advanced and non volatile storage

Motivation

> Analysis only with a web browser
>
> - Available everywhere and at anytime

> Easy to use (but powerful)
>
> - No local installation and configuration needed

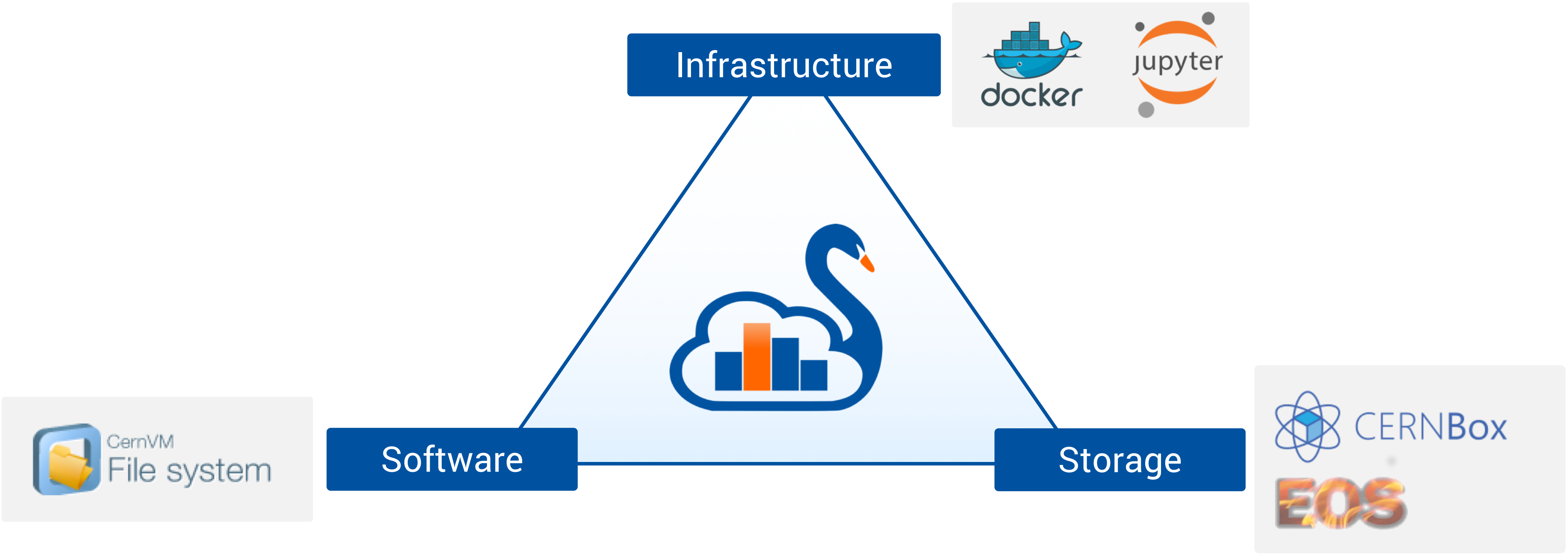
> Create easily sharable scientific results: plots, data, code
>
> - Storage is crucial: mass & synchronized

> Integration with CERN resources
>
> - Access software, user/experiments data, mass processing power

> Integration with other analysis ecosystems : ROOT, R, Python, ...

SWAN

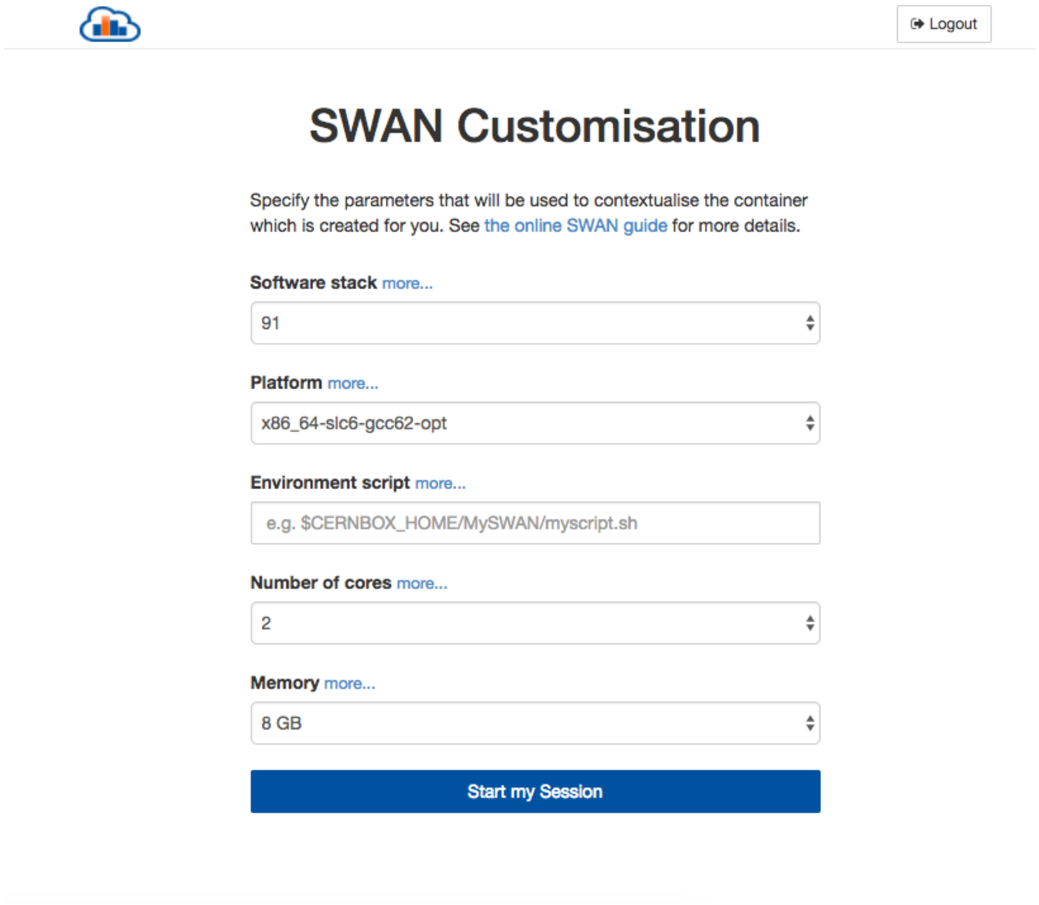
Integrating services



Jupyter - The Notebook as Interface

- A web-based interactive interface and platform that combines code, equations, text and visualisations
  - Ideal for sharing/collaboration
- Many supported languages (kernels): Python, C++, Haskell, Julia, R …
- Very well received Project with major contributions and implementations from big names (IBM, Google,...)
- … In a nutshell: an “interactive shell opened within the browser”

Jupyter - The Notebook as Interface



SWAN Customisation

Specify the parameters that will be used to contextualise the container which is created for you. See the online SWAN guide for more details.

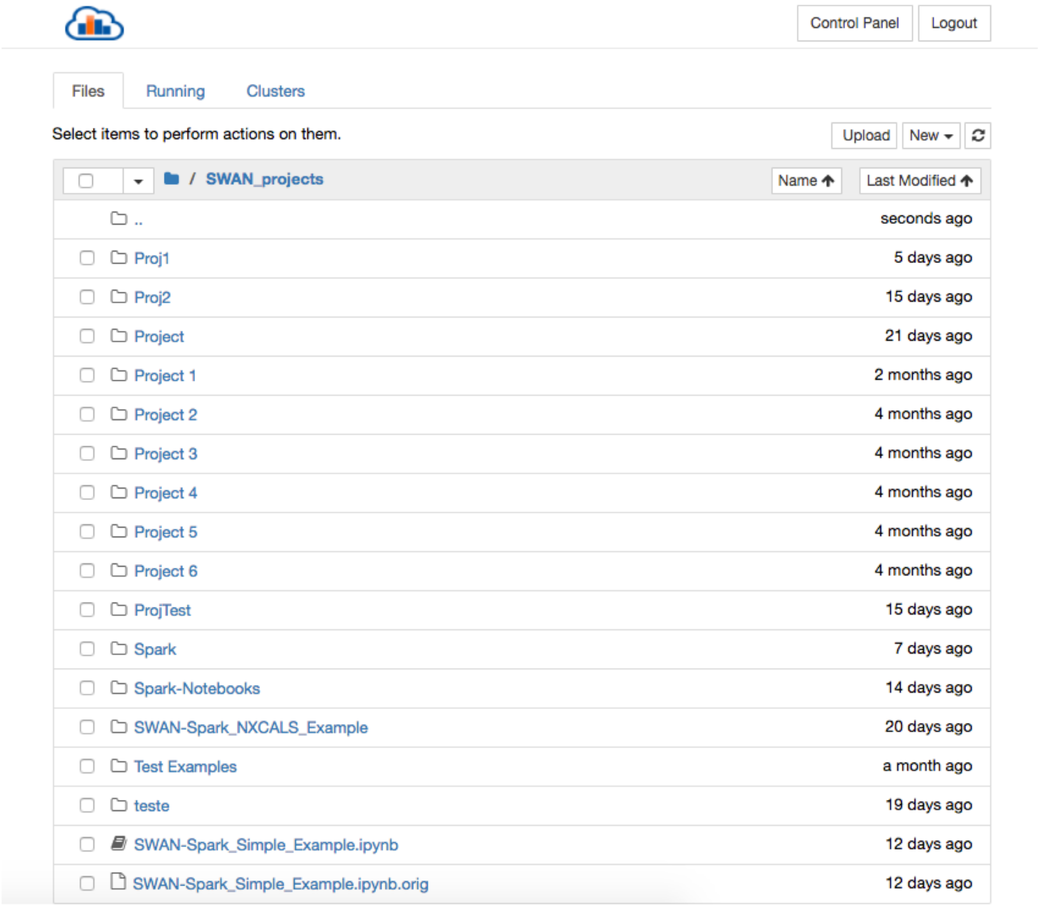
Software stack more...
91

Platform more...
x86_64-slc6-gcc62-opt

Environment script more...
e.g. $CERNBOX_HOME/MySWAN/myscript.sh

Number of cores more...
2

Memory more...
8 GB



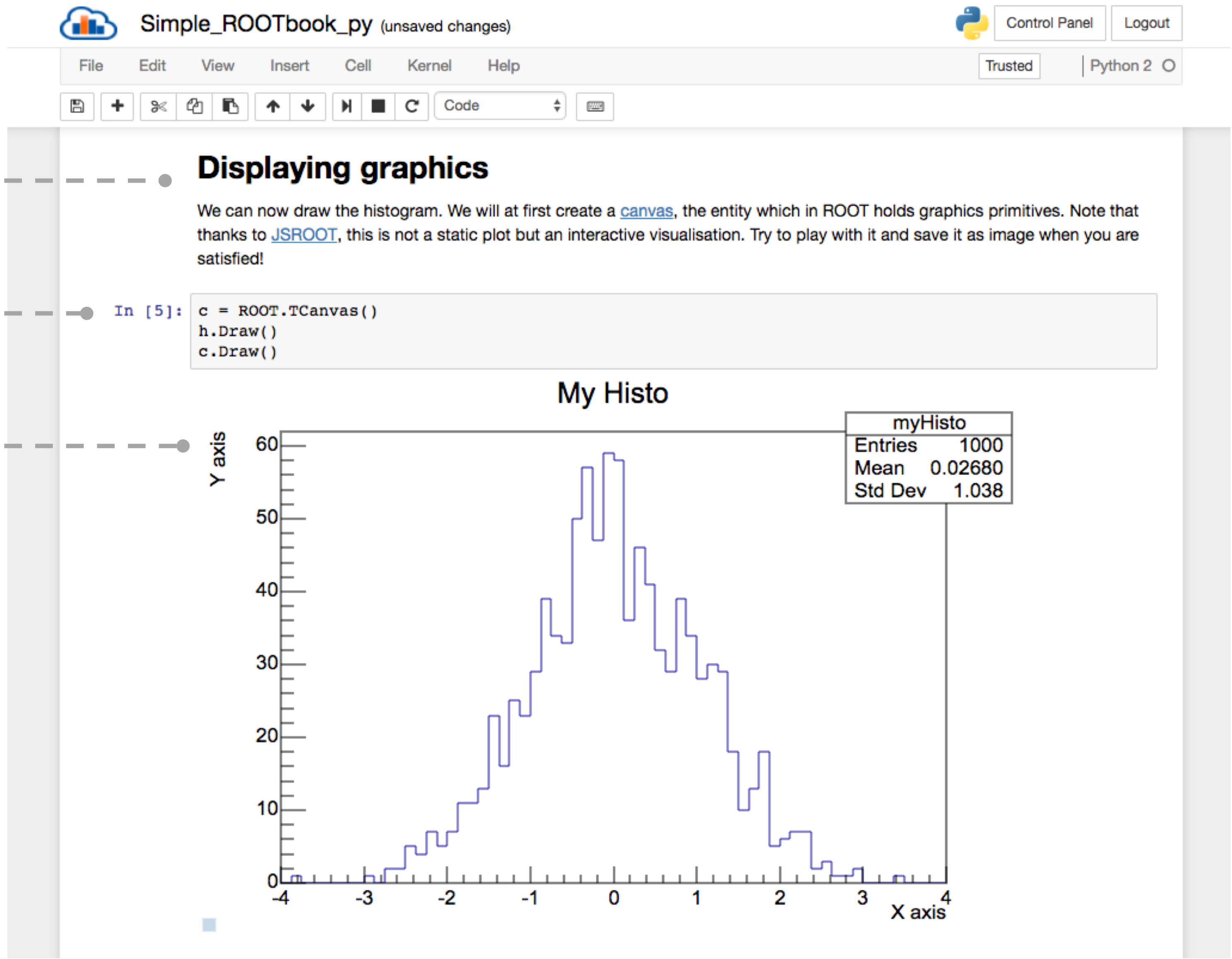
Select items to perform actions on them.

/ SWAN_projects	Last Modified
..	seconds ago
Proj1	5 days ago
Proj2	15 days ago
Project	21 days ago
Project 1	2 months ago
Project 2	4 months ago
Project 3	4 months ago
Project 4	4 months ago
Project 5	4 months ago
Project 6	4 months ago
ProjTest	15 days ago
Spark	7 days ago
Spark-Notebooks	14 days ago
SWAN-Spark_NXCALS_Example	20 days ago
Test Examples	a month ago
teste	19 days ago
SWAN-Spark_Simple_Example.ipynb	12 days ago
SWAN-Spark_Simple_Example.ipynb.orig	12 days ago

Text

Code

Graphics



Displaying graphics

We can now draw the histogram. We will at first create a canvas, the entity which in ROOT holds graphics primitives. Note that thanks to JSROOT, this is not a static plot but an interactive visualisation. Try to play with it and save it as image when you are satisfied!

In [5]:

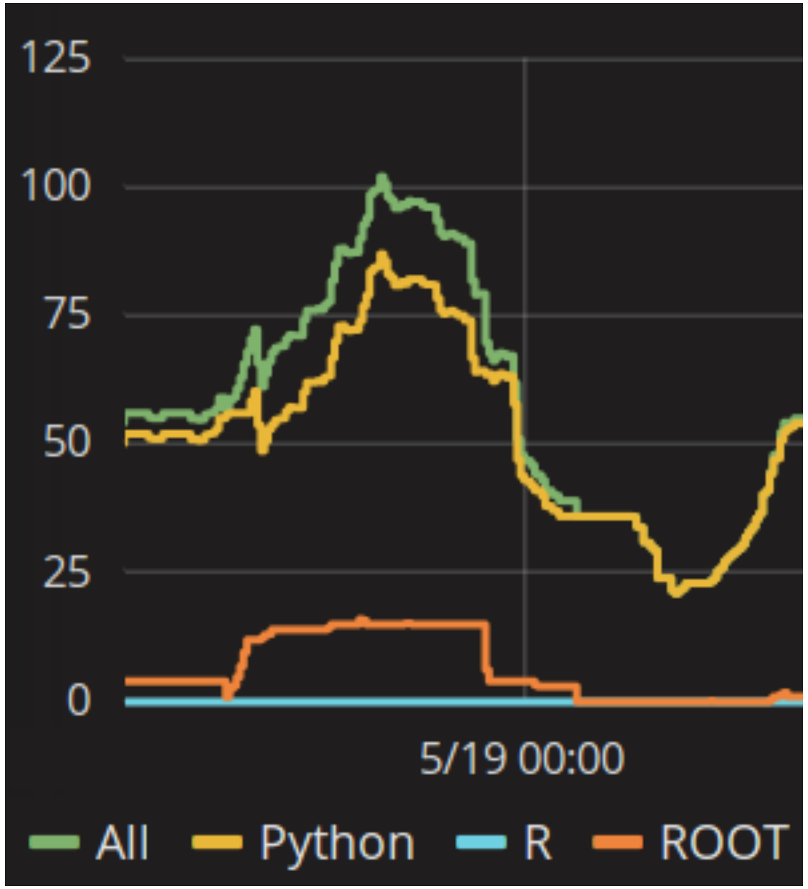
```
c = ROOT.TCanvas()
h.Draw()
c.Draw()
```

Jupyter - The Notebook as Interface

> Very useful for some use cases
  - Final steps of an analysis
  - Exploration
  - Teaching, documentation
  - Reproducibility

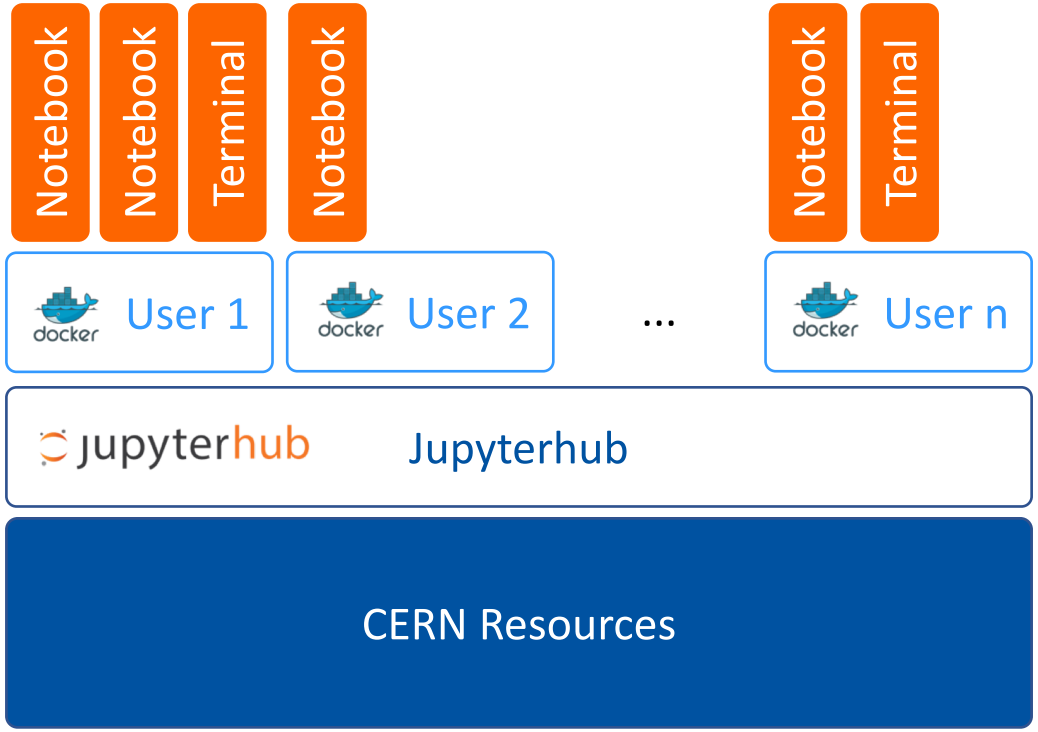
> Interactive, usually lightweight computations

> Languages
  - Not restricted to any
  - Python (2&3) clearly dominating in SWAN

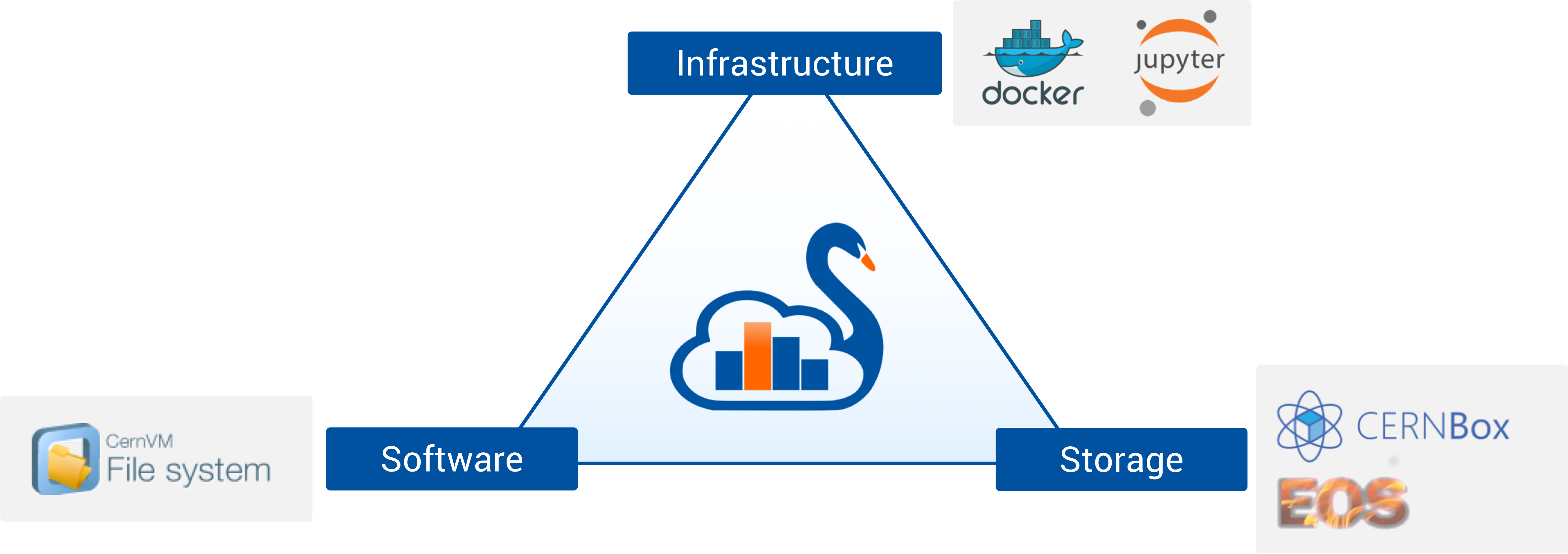


Integrating Jupyter

- Jupyterhub to allow multiple Jupyter instances
- User sessions spawned as Docker containers
  - To guarantee that resources allocated to users are honoured
  - To isolate users work



Integrating services



Infrastructure

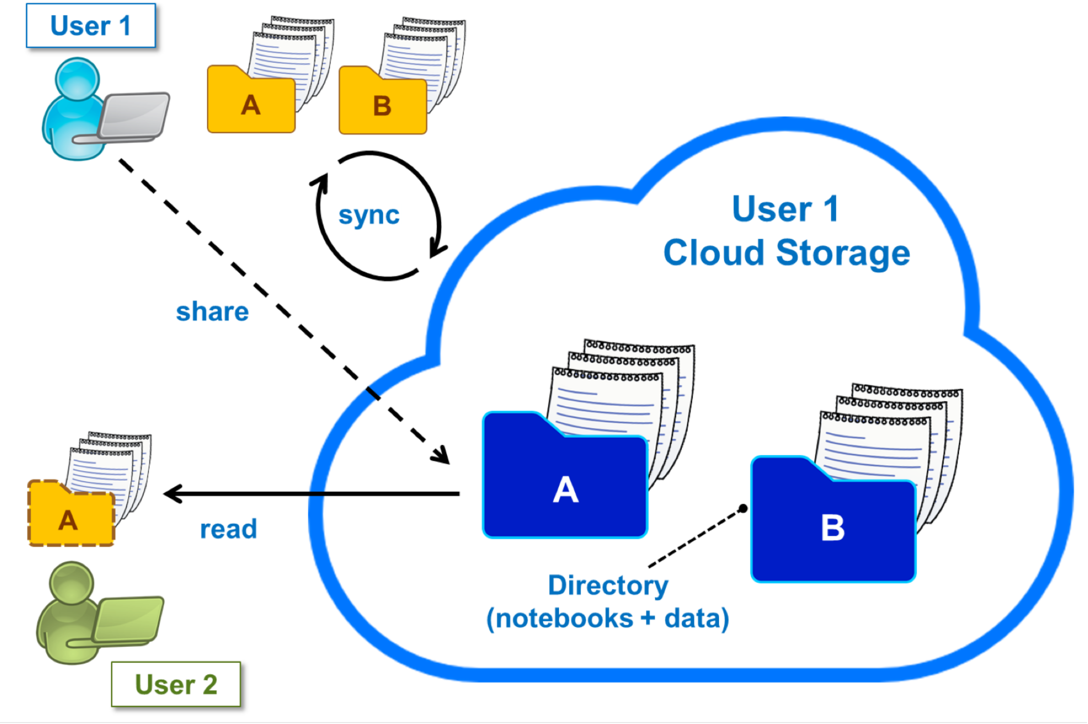
Software

Storage

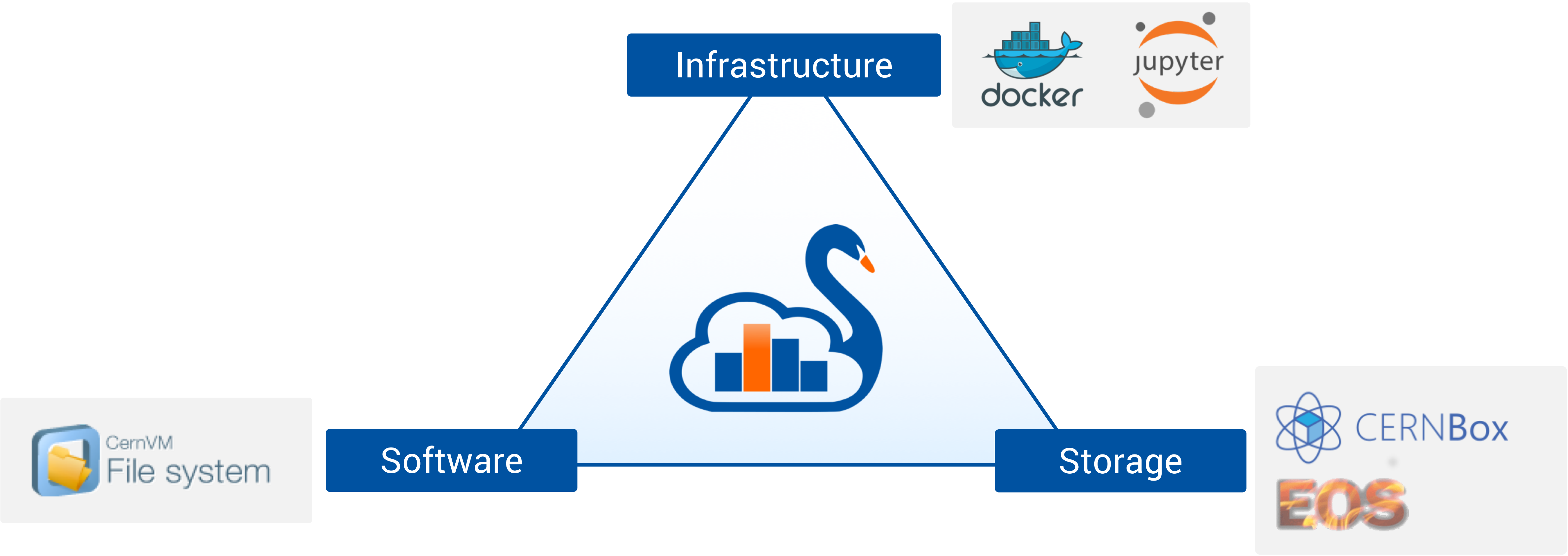
Storage

> Uses EOS mass storage system
  - All experiment data potentially available

> User personal space, synchronized through CERNBox
  - All files synced across devices, the cloud and other users



Integrating services



Infrastructure

Software

Storage

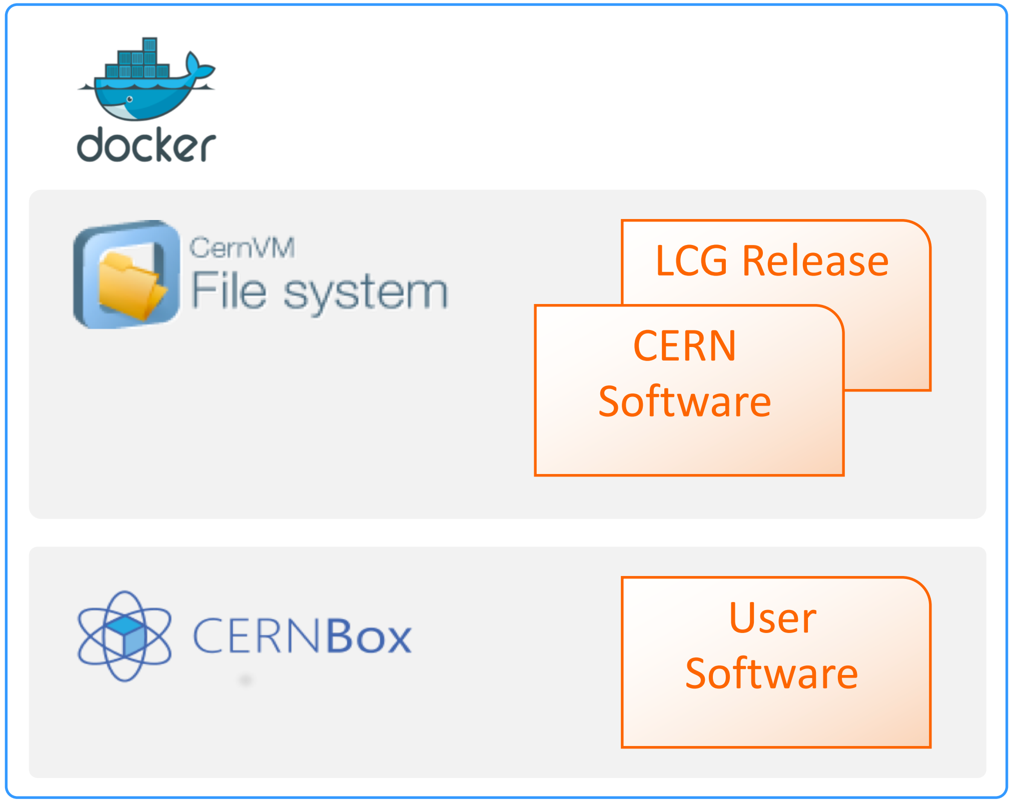
Software

> Software distributed through CVMFS as ”LCG releases”
- A release packs a series of compatible packages

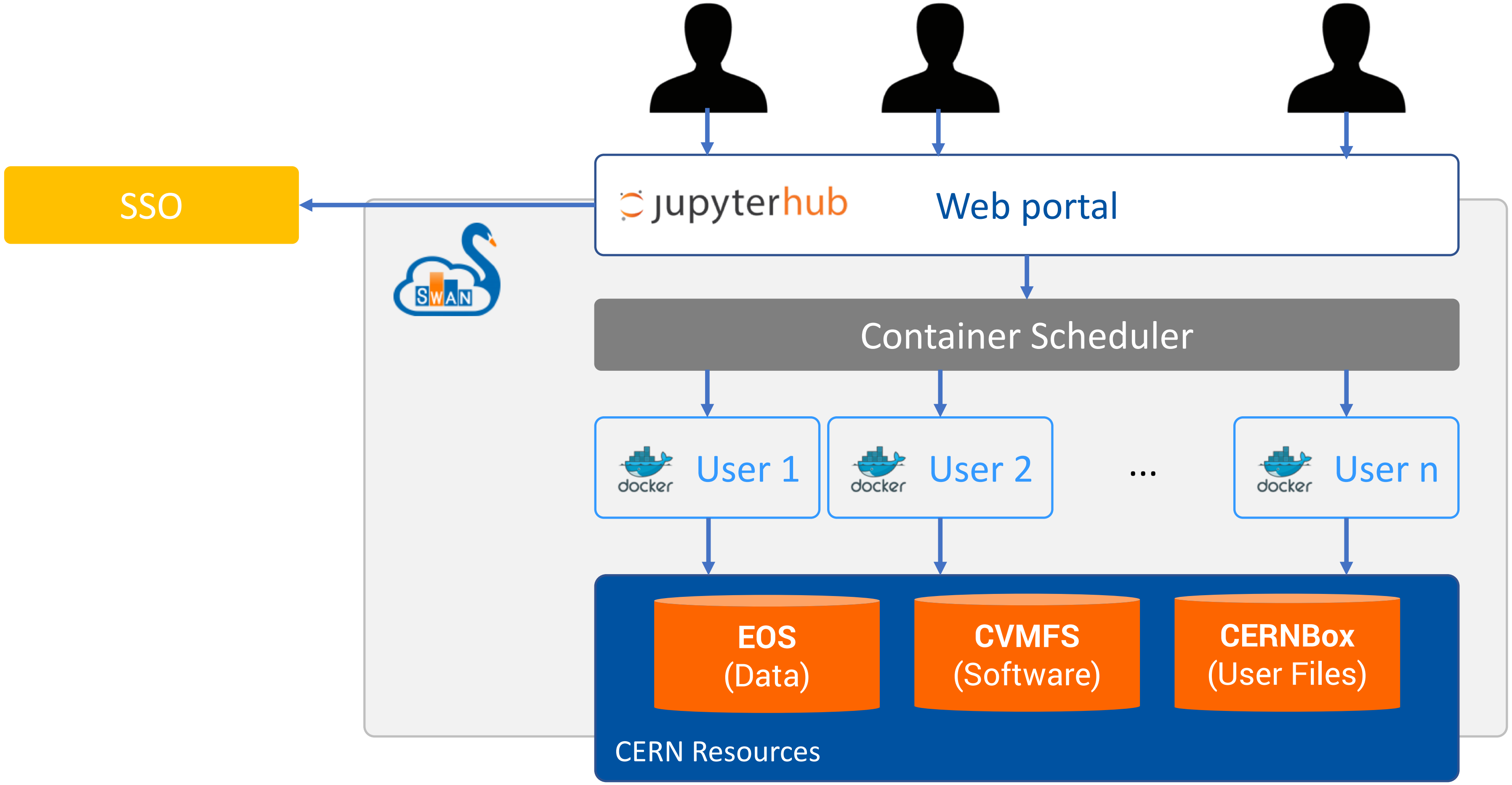
> CVMFS also used by experiments to distribute software
- Software used by researchers is available

> Multiple languages available
- C++ (ROOT), Python, R

> Possibility to install other libraries in user local storage



Architecture

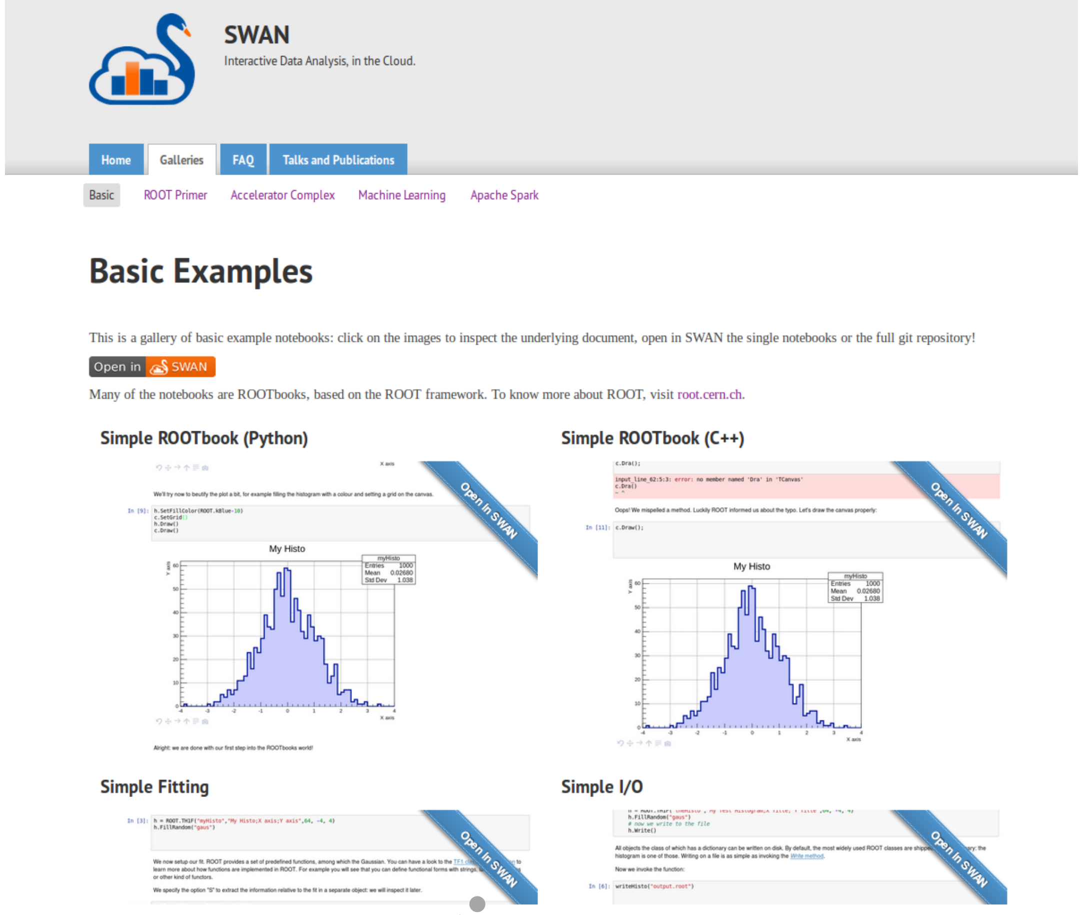


Why SWAN matters

SWAN user community

- SWAN development is guided by our user community
  - New features (libs, kernels, ...) are requested by users from their real usage needs
- Gallery of examples
  - Made in collaboration with our users

Example notebooks at **swan.web.cern.ch**



Basic Examples

This is a gallery of basic example notebooks: click on the images to inspect the underlying document, open in SWAN the single notebooks or the full git repository!

Many of the notebooks are ROOTbooks, based on the ROOT framework. To know more about ROOT, visit root.cern.ch.

Simple ROOTbook (Python)

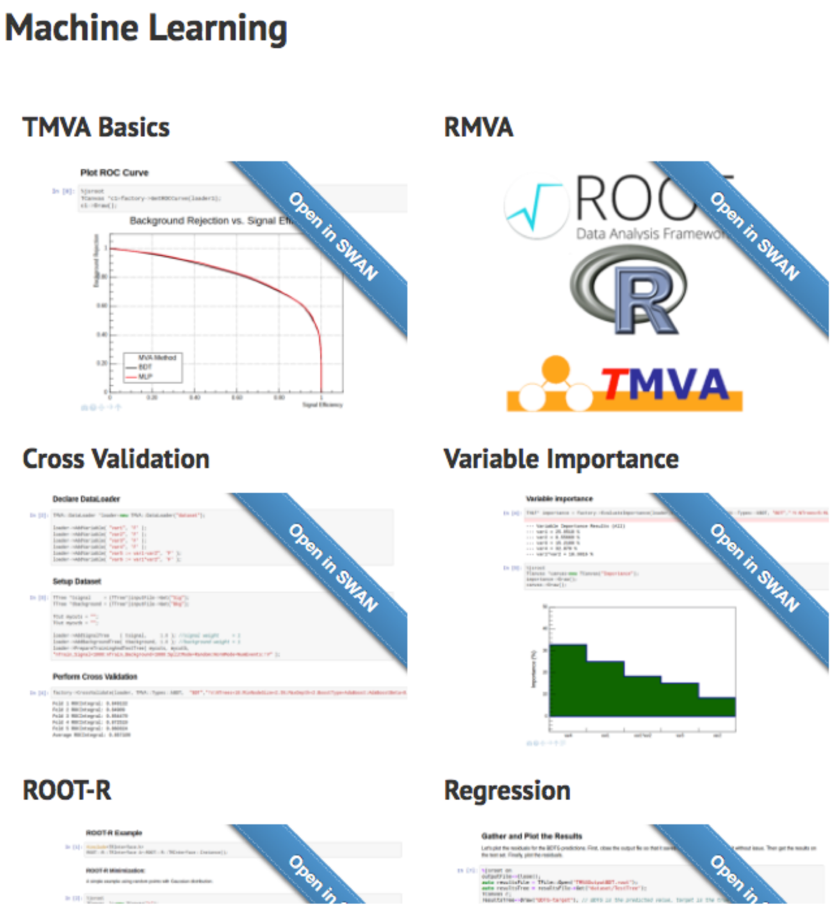
Simple ROOTbook (C++)

Simple Fitting

Simple I/O

Access with only a click

SWAN user community



Machine Learning

TMVA Basics

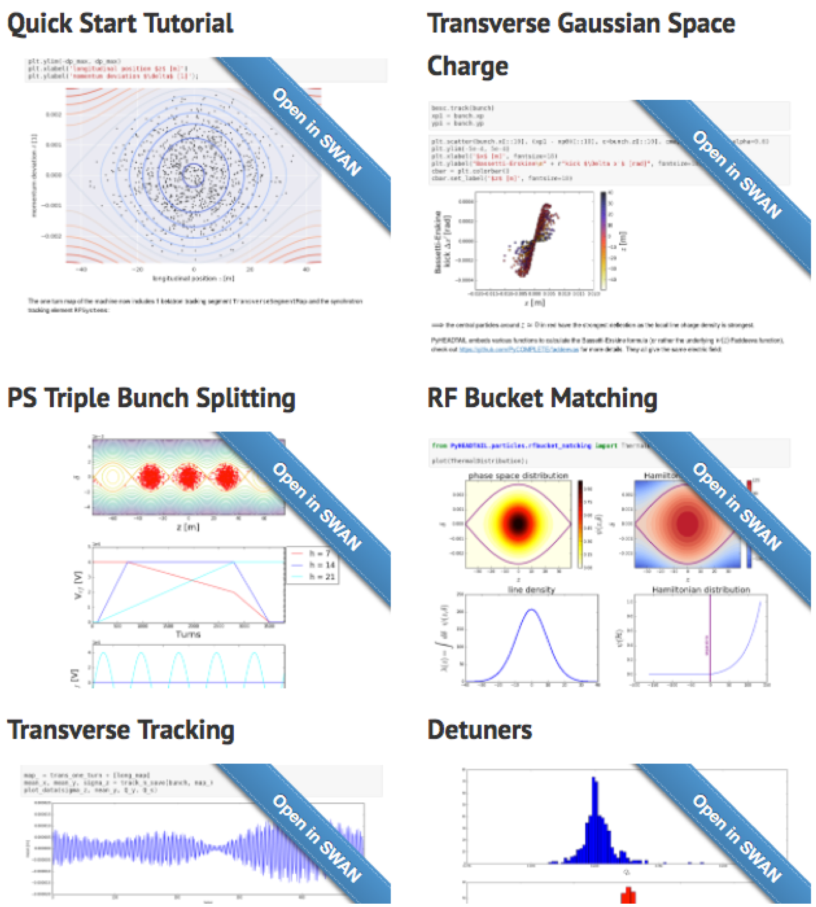
RMVA

Cross Validation

Variable Importance

ROOT-R

Regression



Quick Start Tutorial

Transverse Gaussian Space Charge

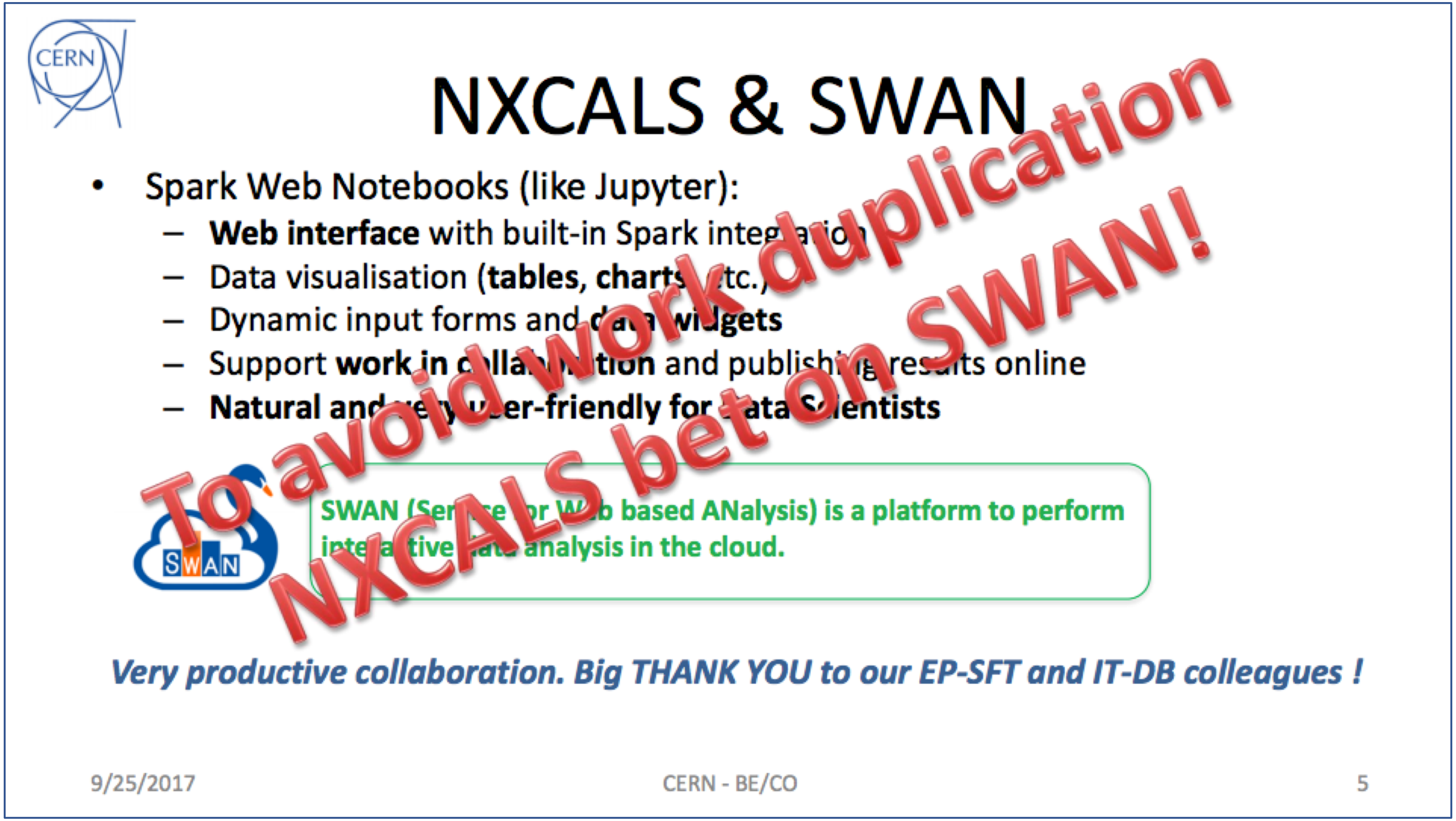
PS Triple Bunch Splitting

RF Bucket Matching

Transverse Tracking

Detuners

Almost 50 notebooks in 7 categories



NXCALS & SWAN

- Spark Web Notebooks (like Jupyter):
  - **Web interface** with built-in Spark integration
  - Data visualisation (**tables, charts** etc.)
  - Dynamic input forms and **data widgets**
  - Support **work in collaboration** and publishing results online
  - **Natural and very user-friendly for Data Scientists**

SWAN (Service for Web based ANalysis) is a platform to perform interactive data analysis in the cloud.

Very productive collaboration. Big THANK YOU to our EP-SFT and IT-DB colleagues !

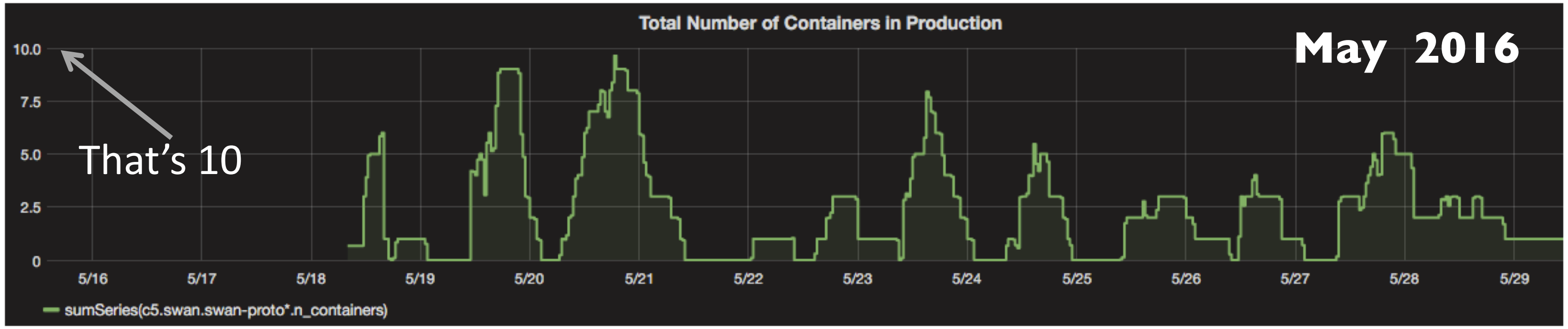
To avoid work duplication

NXCALS bet on SWAN!

9/25/2017 CERN - BE/CO 5

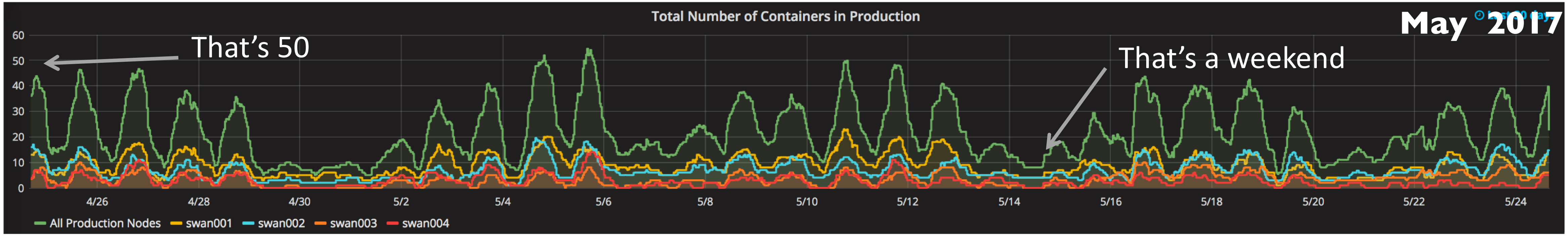
Some users started saying SWAN is fundamental for their work

The numbers



May 2016

That’s 10



May 2017

That’s 50

That’s a weekend

Grand Total since May 2016 (beginning of monitoring):

Number of sessions (containers): **~7k**

Number of notebooks opened: **~14k**

Other collaborations

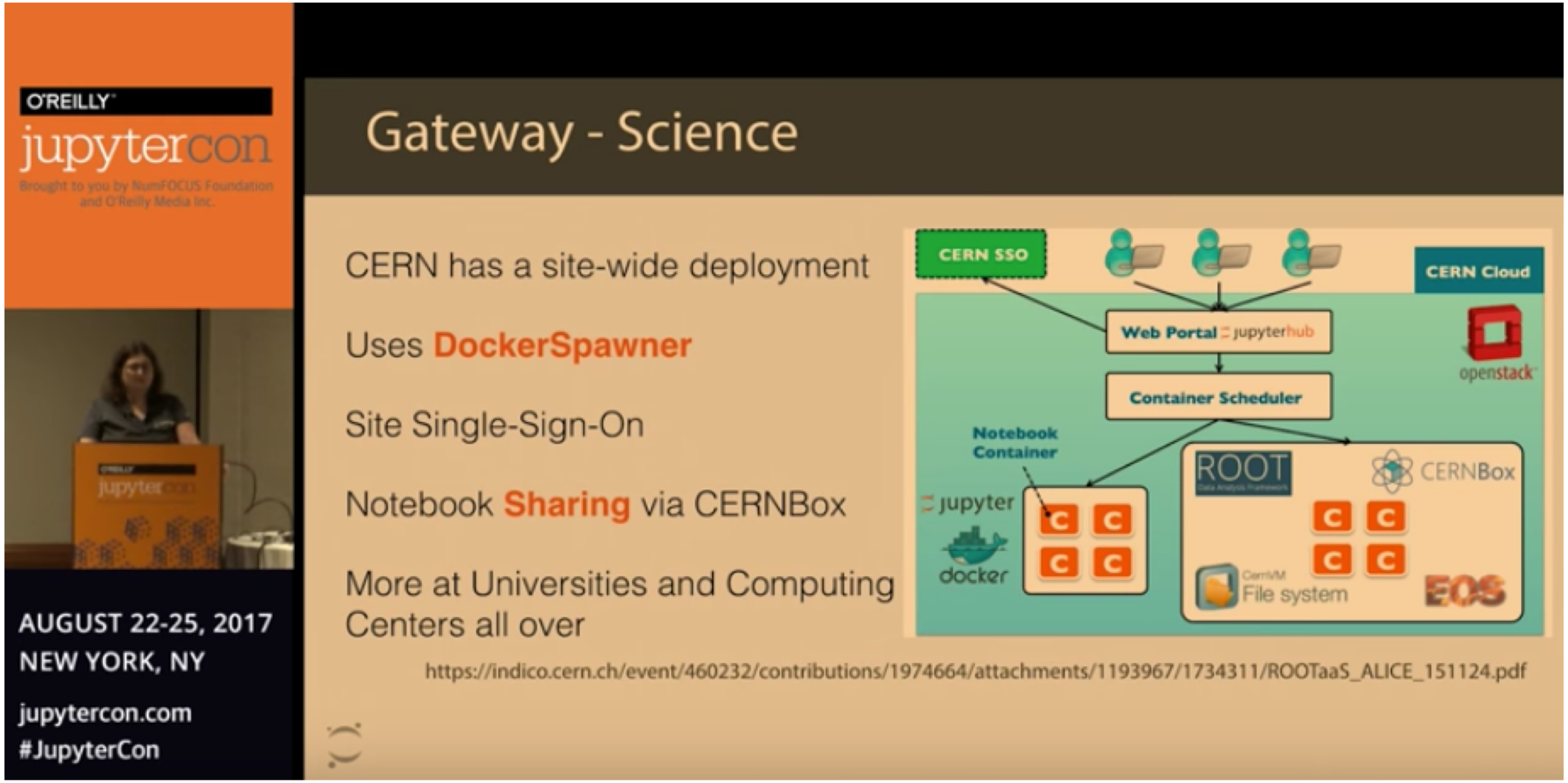
> Building block in UP2University European Project

- Bridge the gap between secondary schools, higher education, and the research domain
- SWAN used by students to learn Physics and other Natural Sciences
- Let the kids use the very same tools & services used by real researchers doing Big Science at CERN

> Will be talked in CS3 Workshop in Krakow

- http://cs3.cyfronet.pl/

SWAN talked outside CERN



O'REILLY

jupytercon

Brought to you by NumFOCUS Foundation and O'Reilly Media Inc.

AUGUST 22-25, 2017
NEW YORK, NY

jupytercon.com
#JupyterCon

Gateway - Science

CERN has a site-wide deployment

Uses **DockerSpawner**

Site Single-Sign-On

Notebook **Sharing** via CERNBox

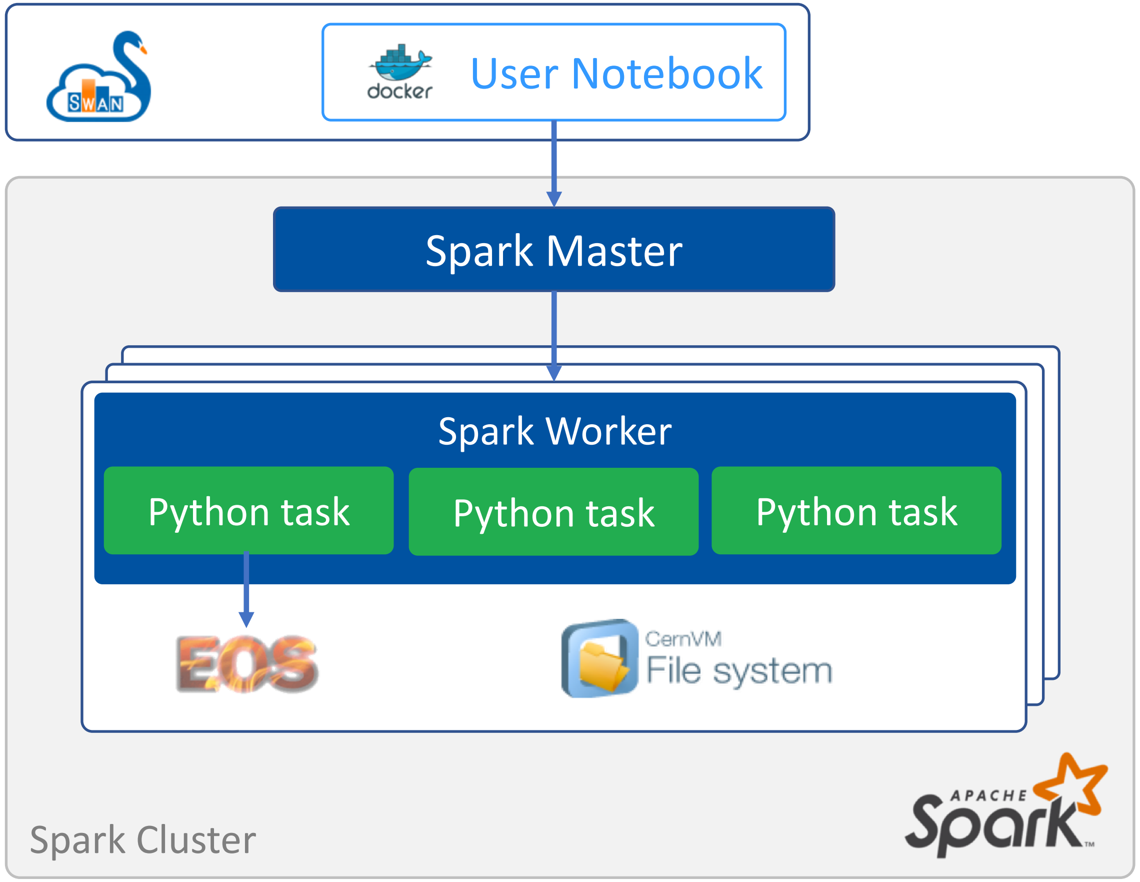
More at Universities and Computing Centers all over

https://indico.cern.ch/event/460232/contributions/1974664/attachments/1193967/1734311/ROOTaaS_ALICE_151124.pdf

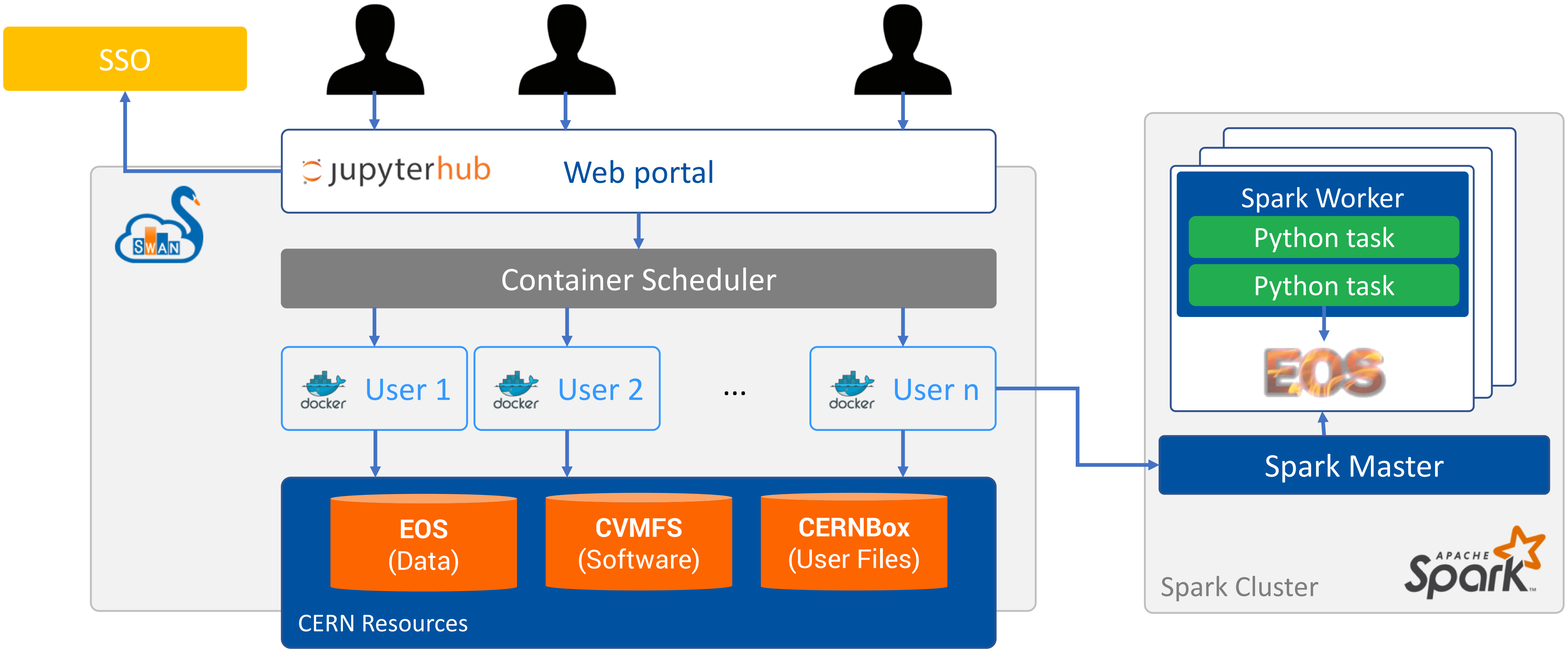
Recent developments

Integration with Spark

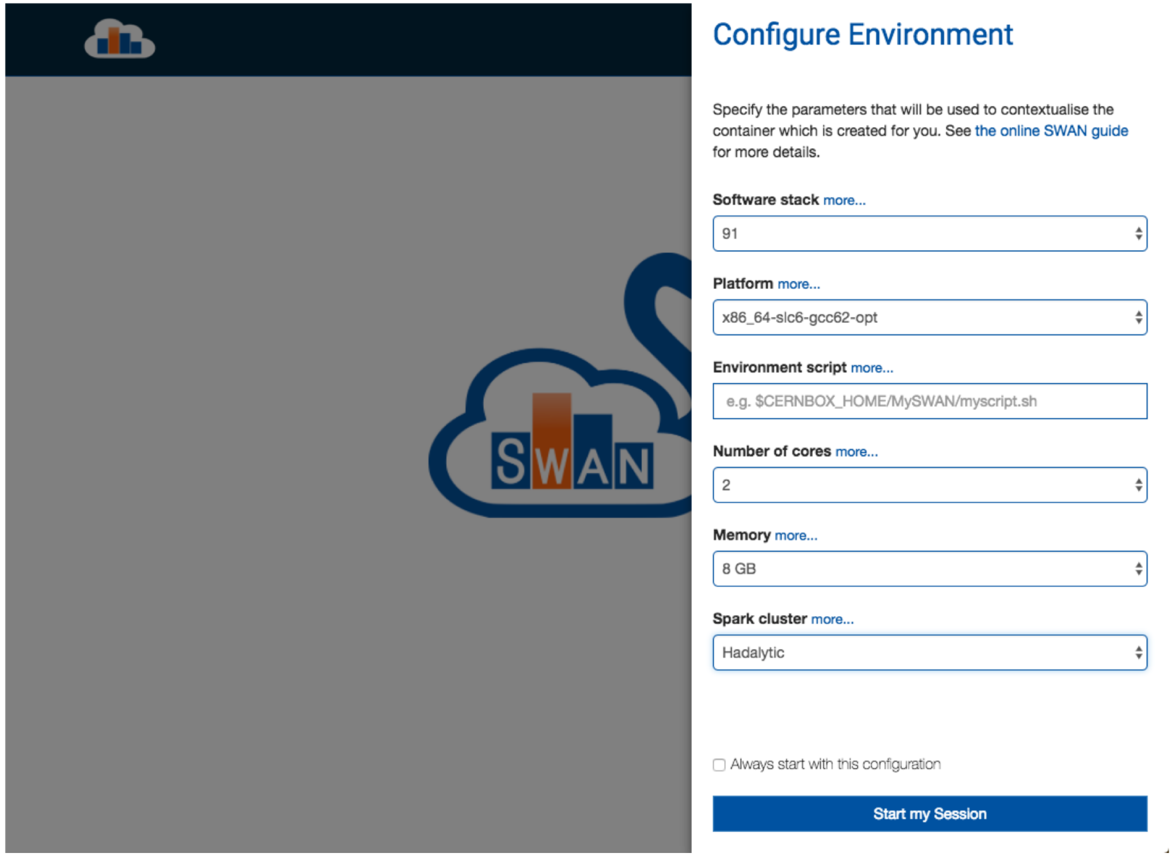
- One of the features requested by the community
  - Team from the Beams department
- Allow users to connect to CERN Spark Clusters to submit jobs
- In collaboration with CERN Database and Storage groups



Architecture



New User Interface



Configure Environment

Specify the parameters that will be used to contextualise the container which is created for you. See the online SWAN guide for more details.

Software stack more...

91

Platform more...

x86_64-slc6-gcc62-opt

Environment script more...

e.g. $CERNBOX_HOME/MySWAN/myscript.sh

Number of cores more...

2

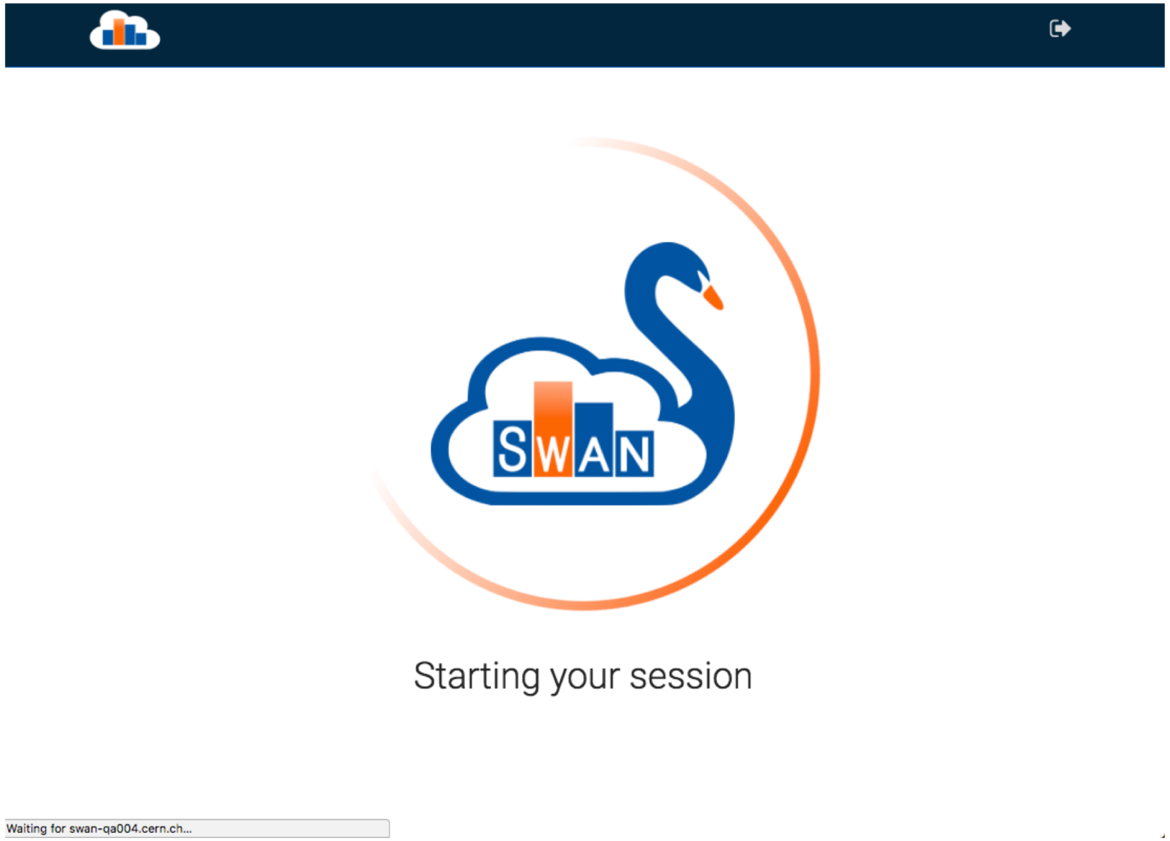
Memory more...

8 GB

Spark cluster more...

Hadalytic

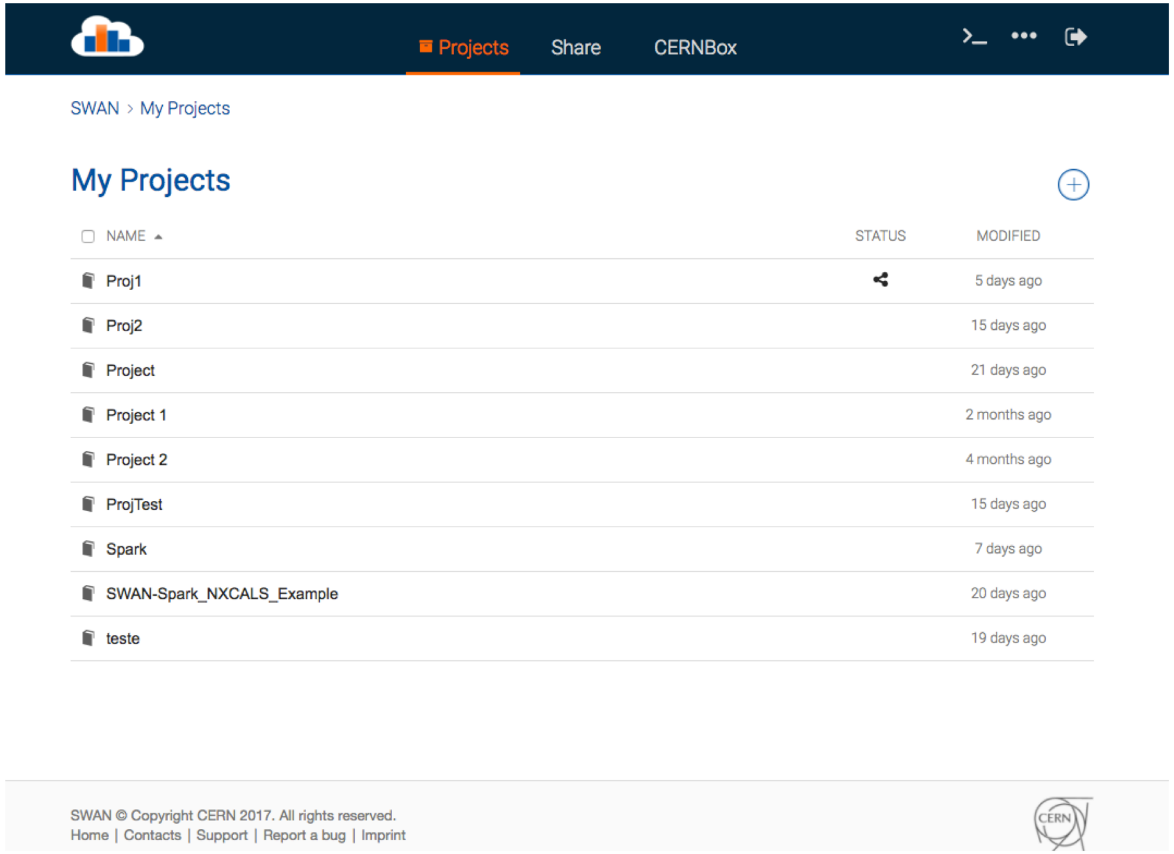
□ Always start with this configuration



Starting your session

Waiting for swan-qa004.cern.ch...

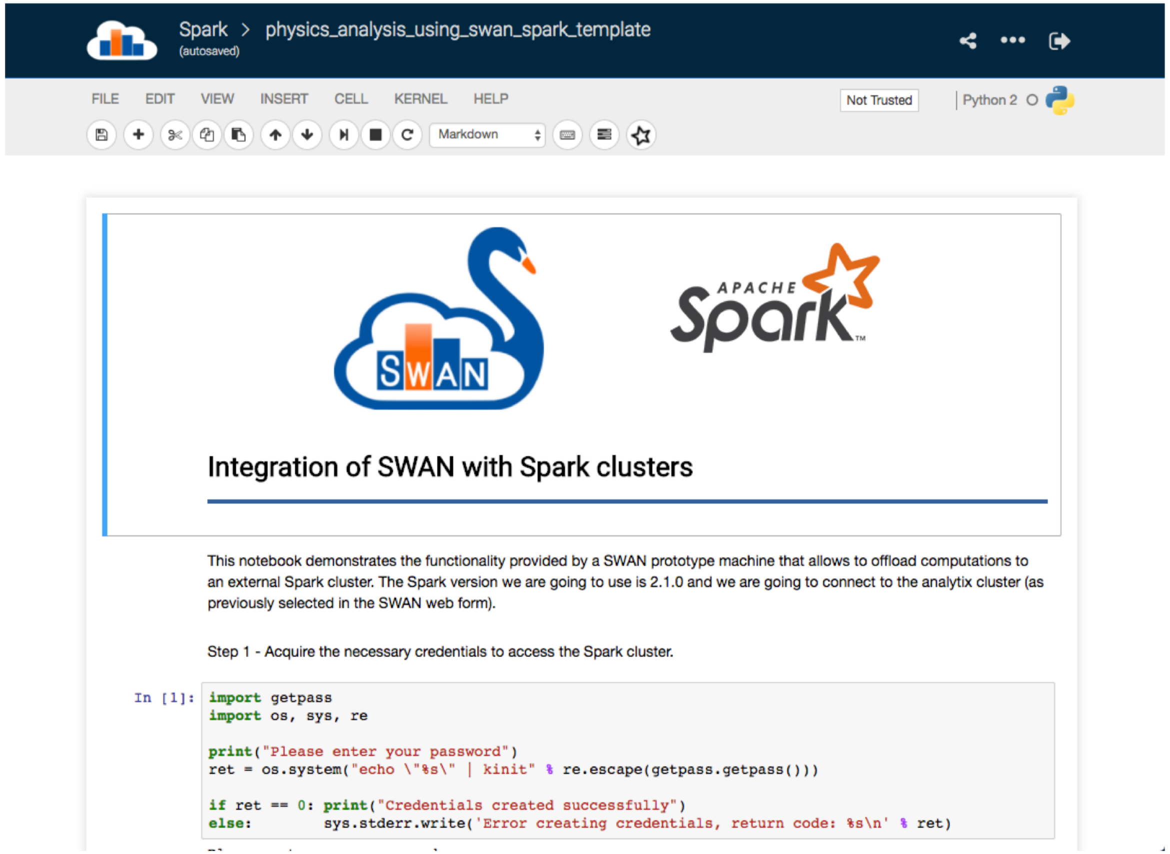
New User Interface



My Projects

□ NAME ▲	STATUS	MODIFIED
Proj1		5 days ago
Proj2		15 days ago
Project		21 days ago
Project 1		2 months ago
Project 2		4 months ago
ProjTest		15 days ago
Spark		7 days ago
SWAN-Spark_NXCALS_Example		20 days ago
teste		19 days ago

SWAN © Copyright CERN 2017. All rights reserved.
Home | Contacts | Support | Report a bug | Imprint



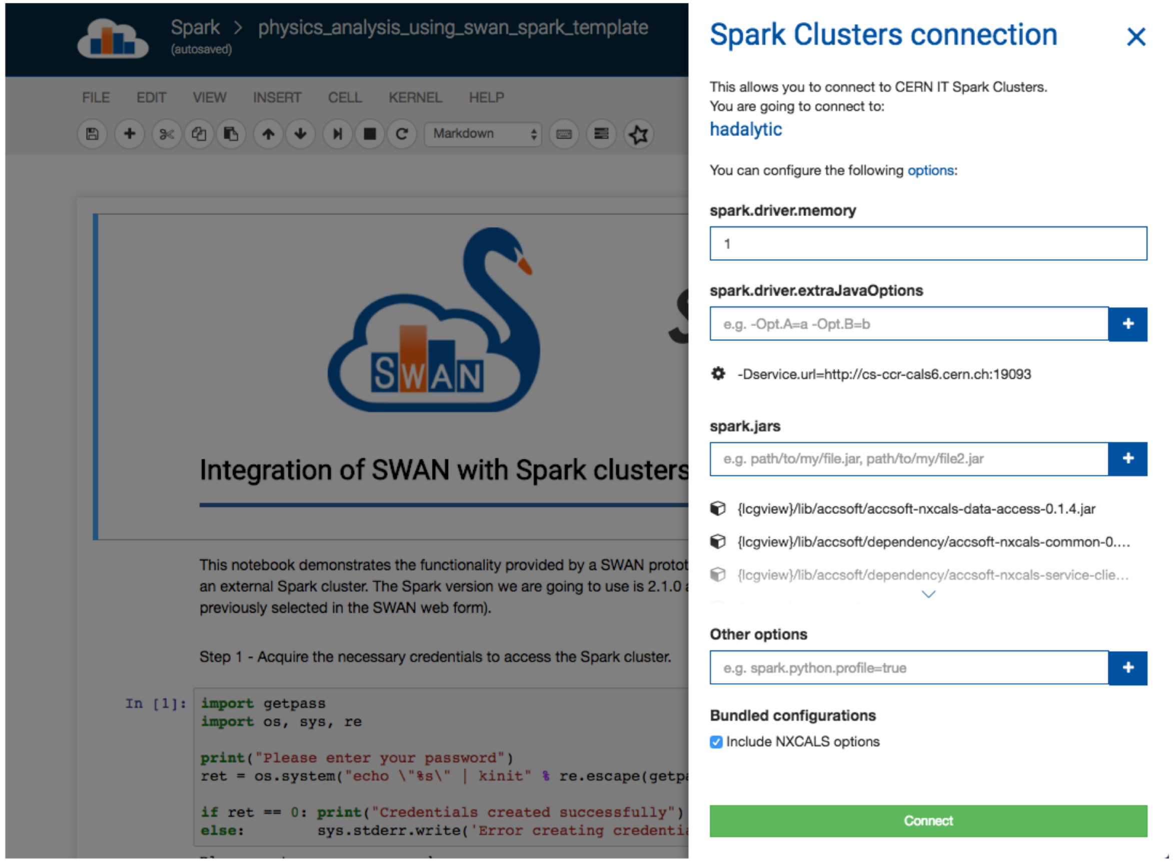
Integration of SWAN with Spark clusters

This notebook demonstrates the functionality provided by a SWAN prototype machine that allows to offload computations to an external Spark cluster. The Spark version we are going to use is 2.1.0 and we are going to connect to the analytix cluster (as previously selected in the SWAN web form).

Step 1 - Acquire the necessary credentials to access the Spark cluster.

```
In [1]: import getpass
        import os, sys, re

        print("Please enter your password")
        ret = os.system("echo \"%s\" | kinit" % re.escape(getpass.getpass()))

        if ret == 0: print("Credentials created successfully")
        else:        sys.stderr.write('Error creating credentials, return code: %s\n' % ret)
```

New User Interface


Spark > physics_analysis_using_swan_spark_template
(autosaved)
FILE EDIT VIEW INSERT CELL KERNEL HELP
Markdown
SWAN

Integration of SWAN with Spark clusters

This notebook demonstrates the functionality provided by a SWAN protot
an external Spark cluster. The Spark version we are going to use is 2.1.0 a
previously selected in the SWAN web form).

Step 1 - Acquire the necessary credentials to access the Spark cluster.

```
In [1]: import getpass
        import os, sys, re

        print("Please enter your password")
        ret = os.system("echo \"%s\" | kinit" % re.escape(getpa

        if ret == 0: print("Credentials created successfully")
        else:        sys.stderr.write('Error creating credenti
```

Spark Clusters connection

This allows you to connect to CERN IT Spark Clusters.
You are going to connect to:
hadalytic

You can configure the following options:

spark.driver.memory

1

spark.driver.extraJavaOptions

e.g. -Opt.A=a -Opt.B=b

-Dservice.url=http://cs-ccr-cals6.cern.ch:19093

spark.jars

e.g. path/to/my/file.jar, path/to/my/file2.jar

{lcgview}/lib/accsoft/accsoft-nxcals-data-access-0.1.4.jar

{lcgview}/lib/accsoft/dependency/accsoft-nxcals-common-0....

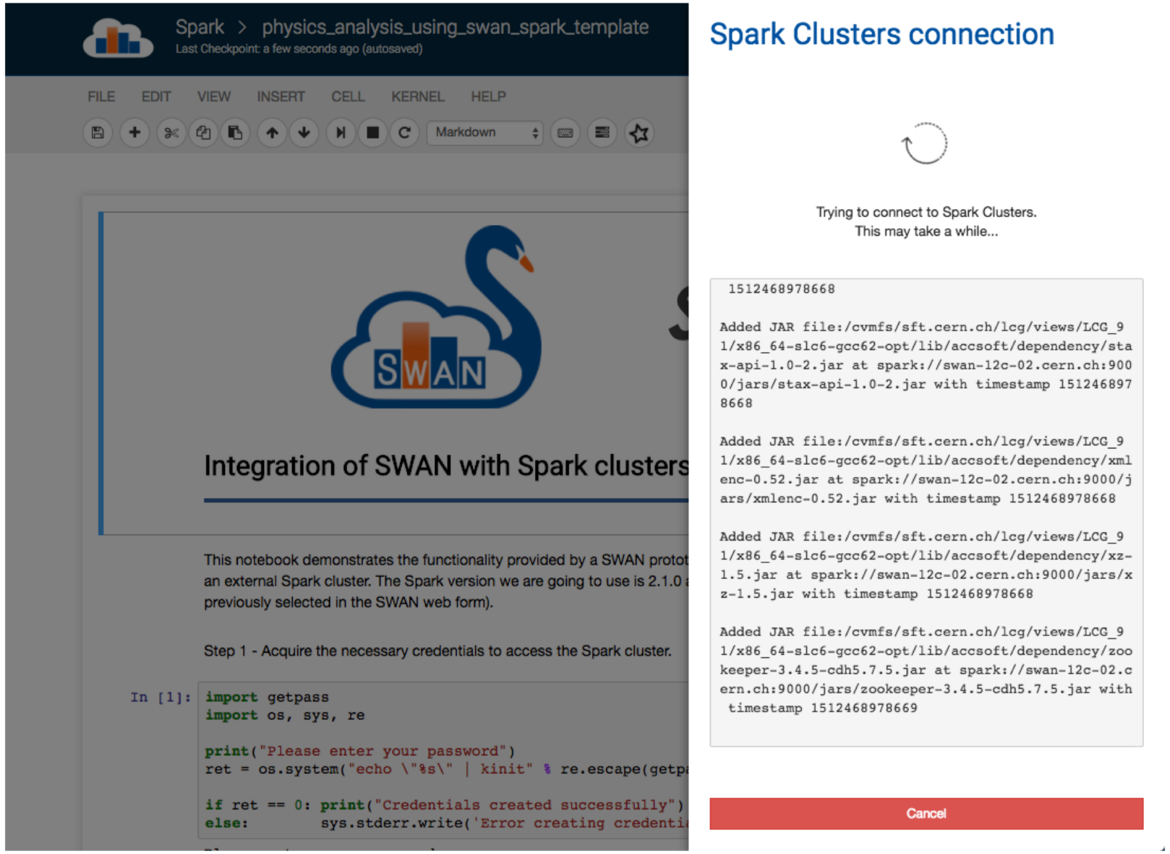
{lcgview}/lib/accsoft/dependency/accsoft-nxcals-service-clie...

Other options

e.g. spark.python.profile=true

Bundled configurations

☑ Include NXCALS options


Spark > physics_analysis_using_swan_spark_template
Last Checkpoint: a few seconds ago (autosaved)
FILE EDIT VIEW INSERT CELL KERNEL HELP
Markdown
SWAN

Integration of SWAN with Spark clusters

This notebook demonstrates the functionality provided by a SWAN protot
an external Spark cluster. The Spark version we are going to use is 2.1.0 a
previously selected in the SWAN web form).

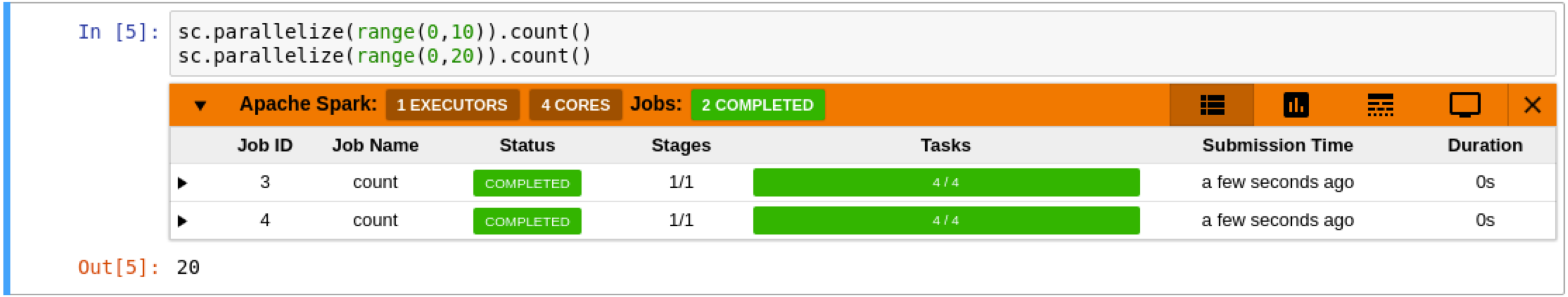
Step 1 - Acquire the necessary credentials to access the Spark cluster.

Spark Clusters connection

Trying to connect to Spark Clusters.
This may take a while...

```
 1512468978668

Added JAR file:/cvmfs/sft.cern.ch/lcg/views/LCG_9
1/x86_64-slc6-gcc62-opt/lib/accsoft/dependency/sta
x-api-1.0-2.jar at spark://swan-12c-02.cern.ch:900
0/jars/stax-api-1.0-2.jar with timestamp 151246897
8668

Added JAR file:/cvmfs/sft.cern.ch/lcg/views/LCG_9
1/x86_64-slc6-gcc62-opt/lib/accsoft/dependency/xml
enc-0.52.jar at spark://swan-12c-02.cern.ch:9000/j
ars/xmlenc-0.52.jar with timestamp 1512468978668

Added JAR file:/cvmfs/sft.cern.ch/lcg/views/LCG_9
1/x86_64-slc6-gcc62-opt/lib/accsoft/dependency/xz-
1.5.jar at spark://swan-12c-02.cern.ch:9000/jars/x
z-1.5.jar with timestamp 1512468978668

Added JAR file:/cvmfs/sft.cern.ch/lcg/views/LCG_9
1/x86_64-slc6-gcc62-opt/lib/accsoft/dependency/zoo
keeper-3.4.5-cdh5.7.5.jar at spark://swan-12c-02.c
ern.ch:9000/jars/zookeeper-3.4.5-cdh5.7.5.jar with
 timestamp 1512468978669
```

New User Interface



```
In [5]: sc.parallelize(range(0,10)).count()
        sc.parallelize(range(0,20)).count()
```

▼ Apache Spark: 1 EXECUTORS 4 CORES Jobs: 2 COMPLETED

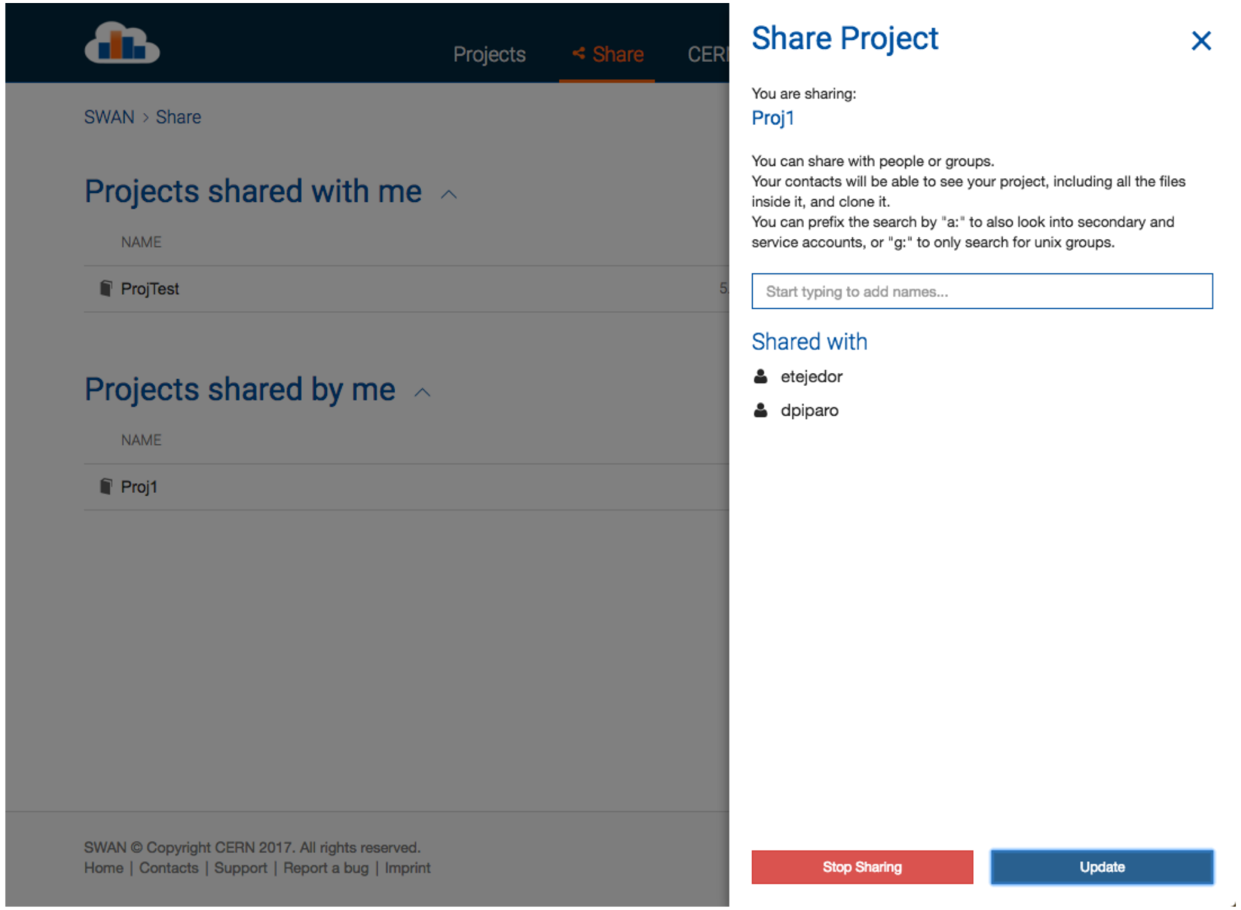
	Job ID	Job Name	Status	Stages	Tasks	Submission Time	Duration
▶	3	count	COMPLETED	1/1	4 / 4	a few seconds ago	0s
▶	4	count	COMPLETED	1/1	4 / 4	a few seconds ago	0s

```
Out[5]: 20
```

Sharing made easy

> Sharing from inside SWAN interface

> Users can share “Projects”
  - Special kind of folder that contains notebooks and other files, like input data



Projects shared with me

ProjTest

Projects shared by me

Proj1

SWAN © Copyright CERN 2017. All rights reserved.
Home | Contacts | Support | Report a bug | Imprint

Share Project

You are sharing:
Proj1

You can share with people or groups.
Your contacts will be able to see your project, including all the files inside it, and clone it.
You can prefix the search by "a:" to also look into secondary and service accounts, or "g:" to only search for unix groups.

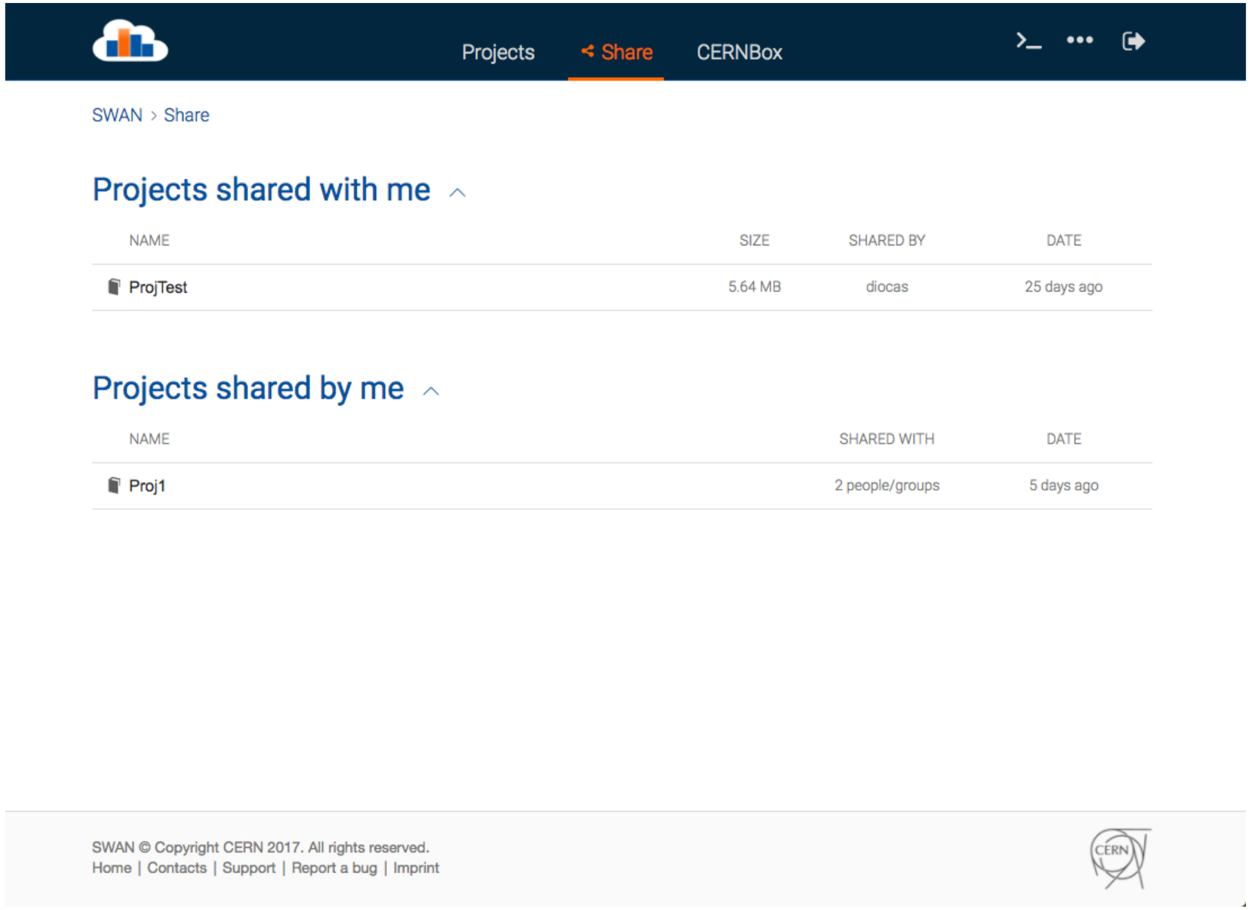
Start typing to add names...

Shared with

- etejedor
- dpiparo

Sharing made easy

> Users can clone a shared Project directly from the interface



SWAN > Share

Projects shared with me

NAME	SIZE	SHARED BY	DATE
ProjTest	5.64 MB	diocas	25 days ago

Projects shared by me

NAME	SHARED WITH	DATE
Proj1	2 people/groups	5 days ago

SWAN © Copyright CERN 2017. All rights reserved.
Home | Contacts | Support | Report a bug | Imprint

New functionalities

- New Octave kernel
- Possible integration with HTCondor

Conclusion

Conclusion

- SWAN is a CERN service that provides Jupyter Notebooks on demand
- SWAN promotes a cloud based analysis model where users can do analysis only with their browser
- SWAN federates CERN services for software, storage and infrastructure so that users can find what they need in the service
- SWAN fosters collaboration and results sharing between scientists
- SWAN is an Interface for Mass Processing Resources (Spark)

SWAN: service for web based analysis

Thank you