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

Mar 13th, 2018

Jupyter notebooks at CERN – EPFL meeting

Introduction

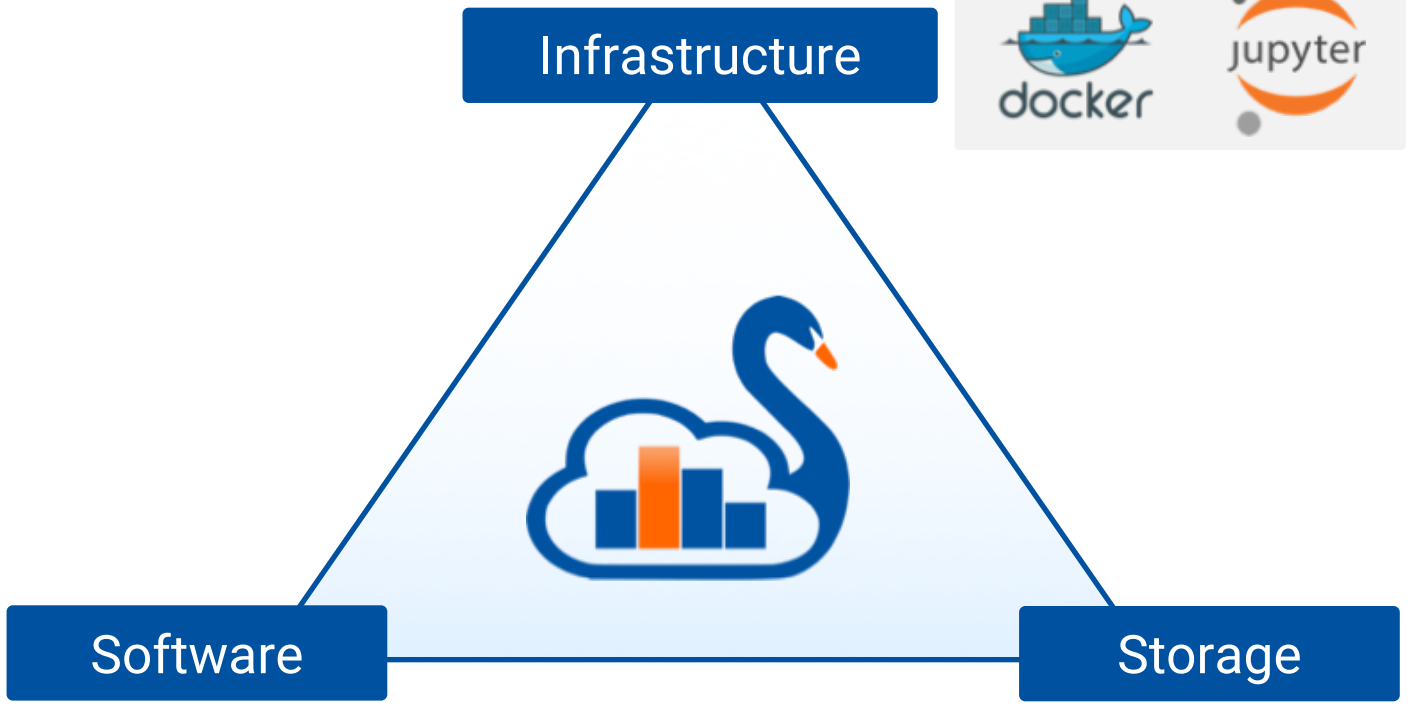
Motivation

- Analysis only with a web browser
  - Available everywhere and at anytime
  - Integrated with other analysis ecosystems : ROOT, R, Python, ...
- Easy to use (but powerful)
  - No local installation and configuration needed
- Create easily sharable scientific results: plots, data, code
- Integration with CERN resources
  - Access software, user/experiments data, mass processing power

CERN

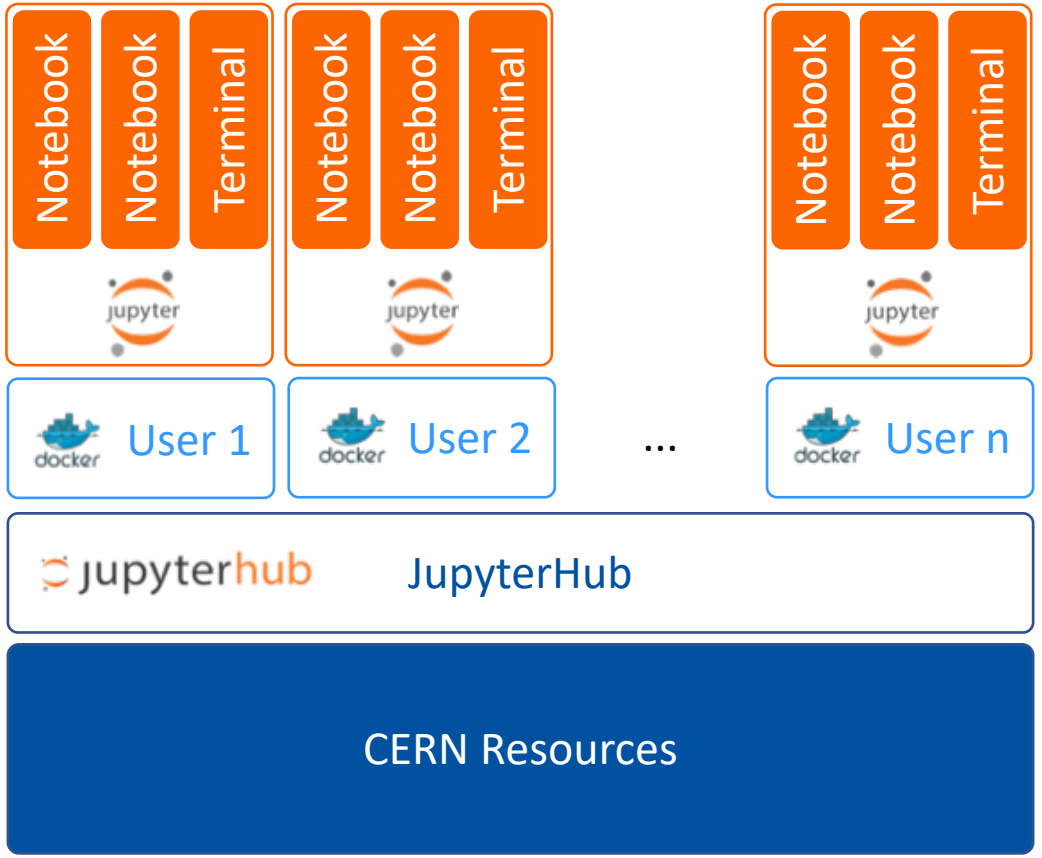
SWAN

Integrating services



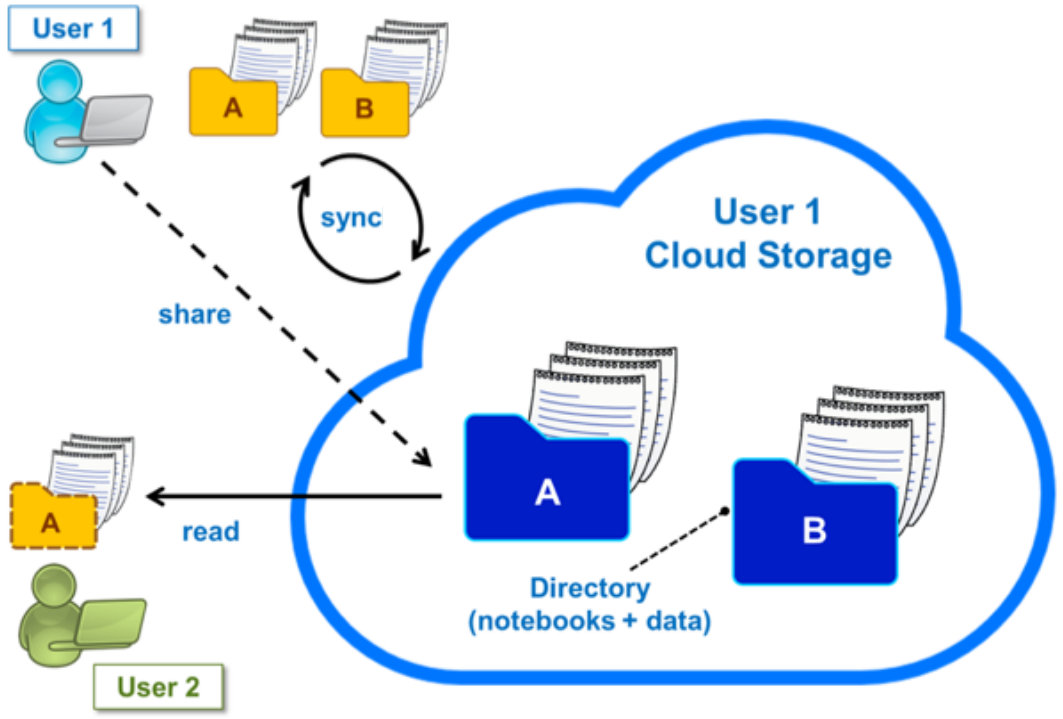
Integrating Jupyter

- 4 kernels
  - Python (2 or 3), ROOT C++, R and Octave
- Configurable environments through user defined scripts
- Jupyterhub to allow multiple Jupyter instances
  - Single instance of Jupyter per user
  - Proxy the container to the outside
- User sessions spawned as Docker containers
  - Enforces resource limits per user
  - Sessions culled after 6h of inactivity



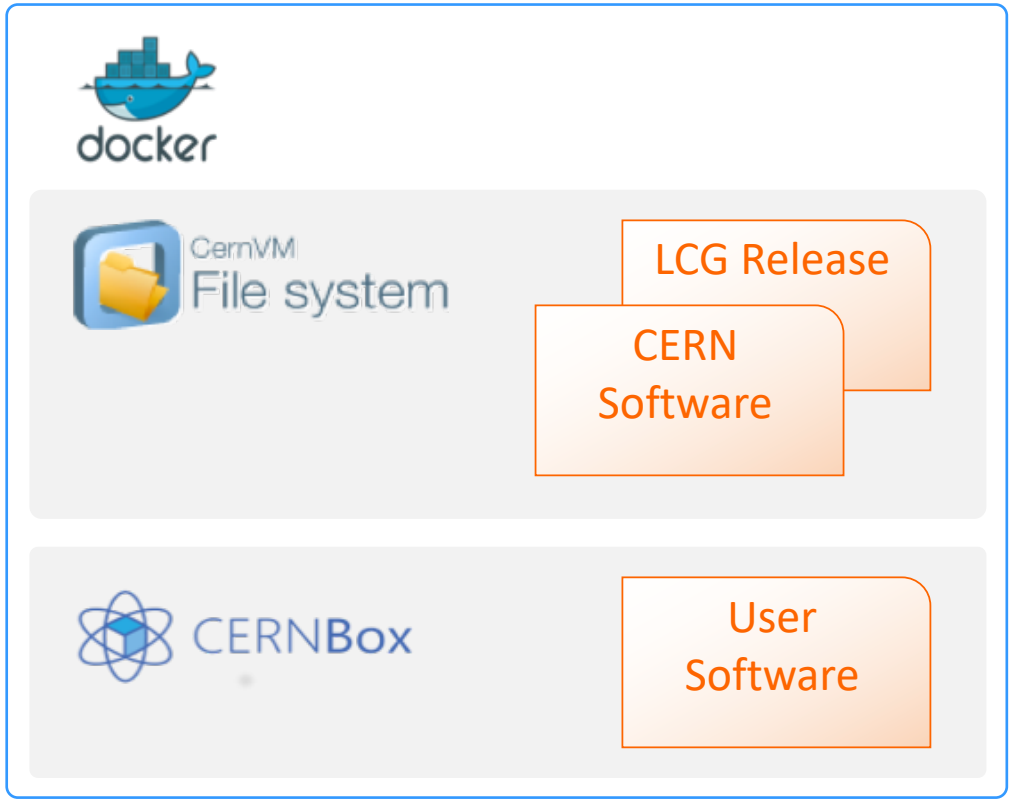
Storage

- Uses EOS mass storage system
  - All experiment data potentially available
  - Installed on host, mounted on container
- User personal space, synchronized through CERNBox
  - All files synced across devices, the cloud and other users
  - Based on OwnCloud

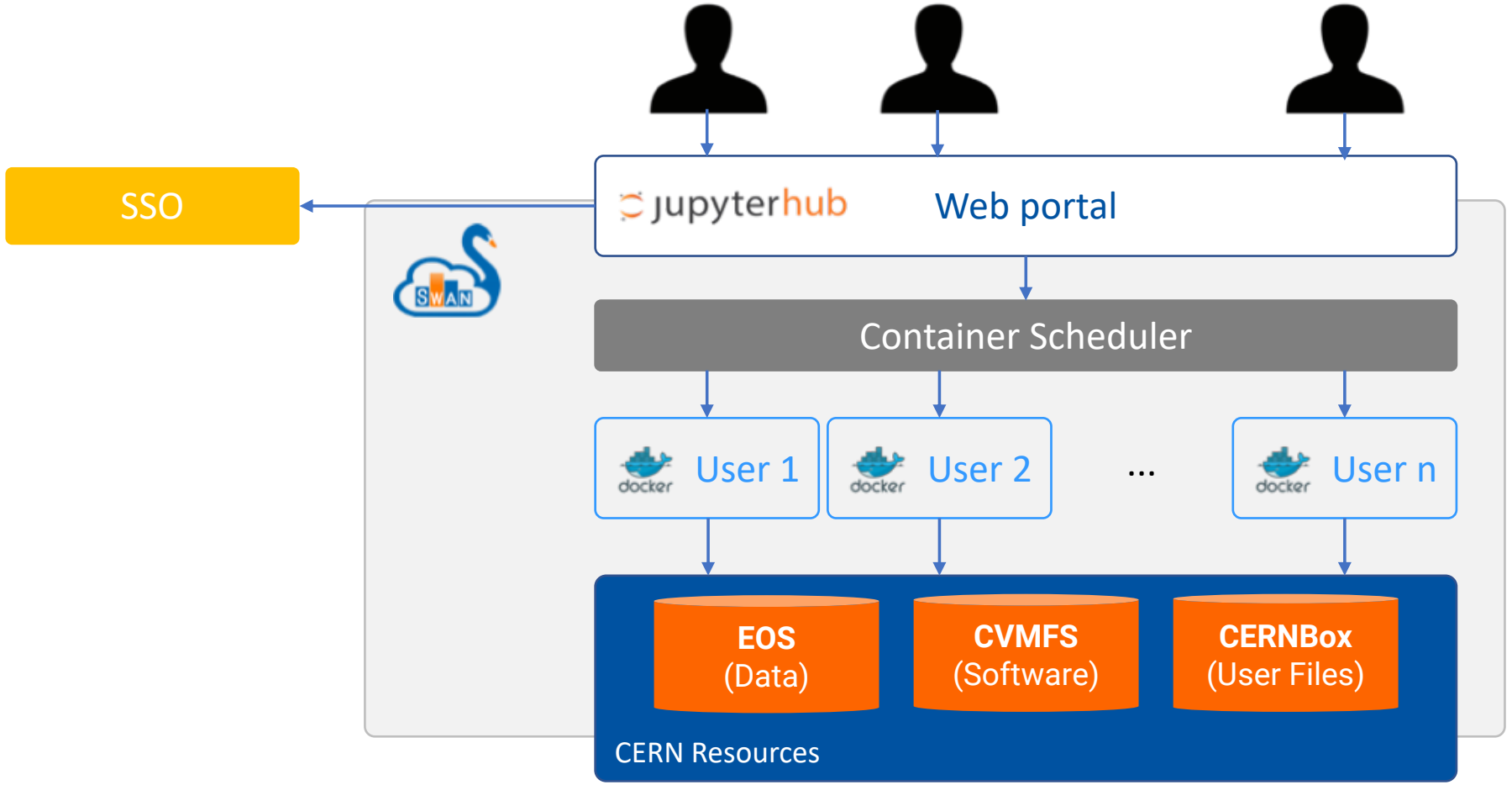


Software

- Software distributed through CVMFS
  - ”LCG Releases” - pack a series of compatible packages
  - Software used by researchers is available
  - Strong caching mechanisms
- Possibility to install other libraries in user storage (CERNBox)



Architecture

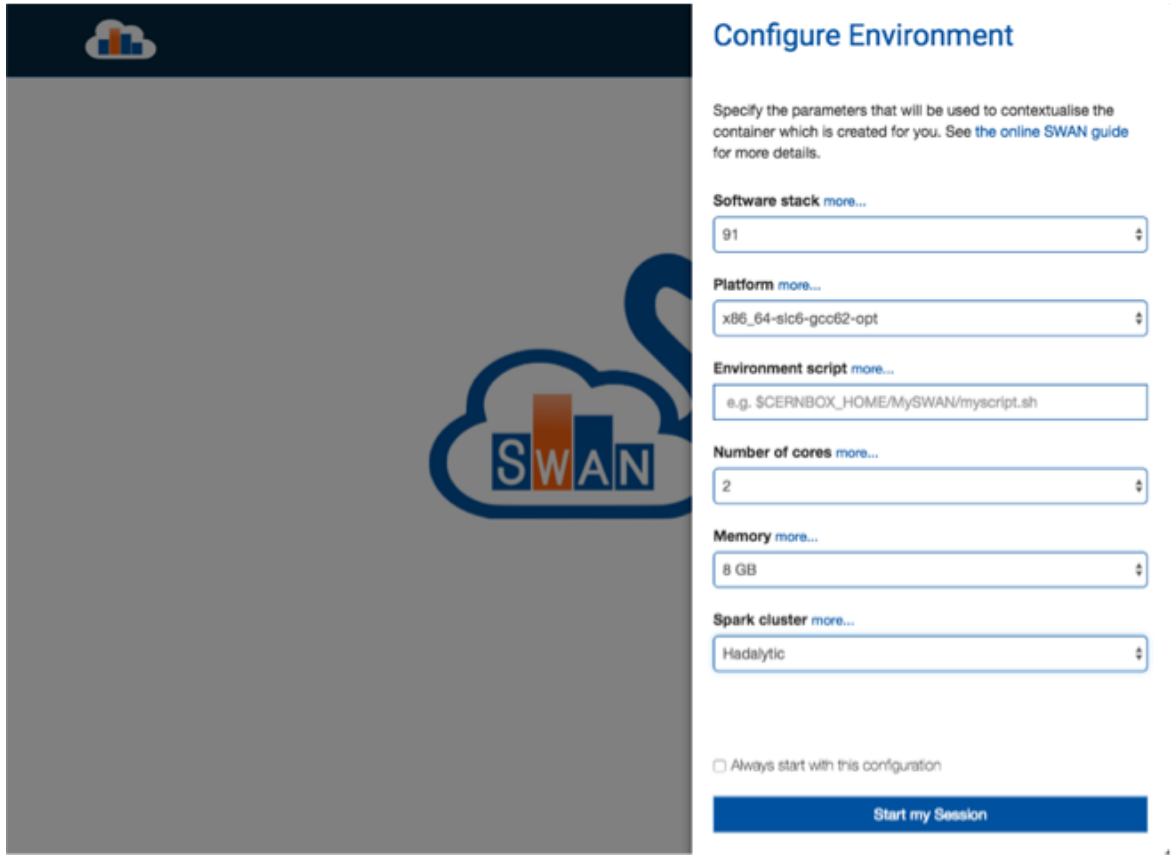


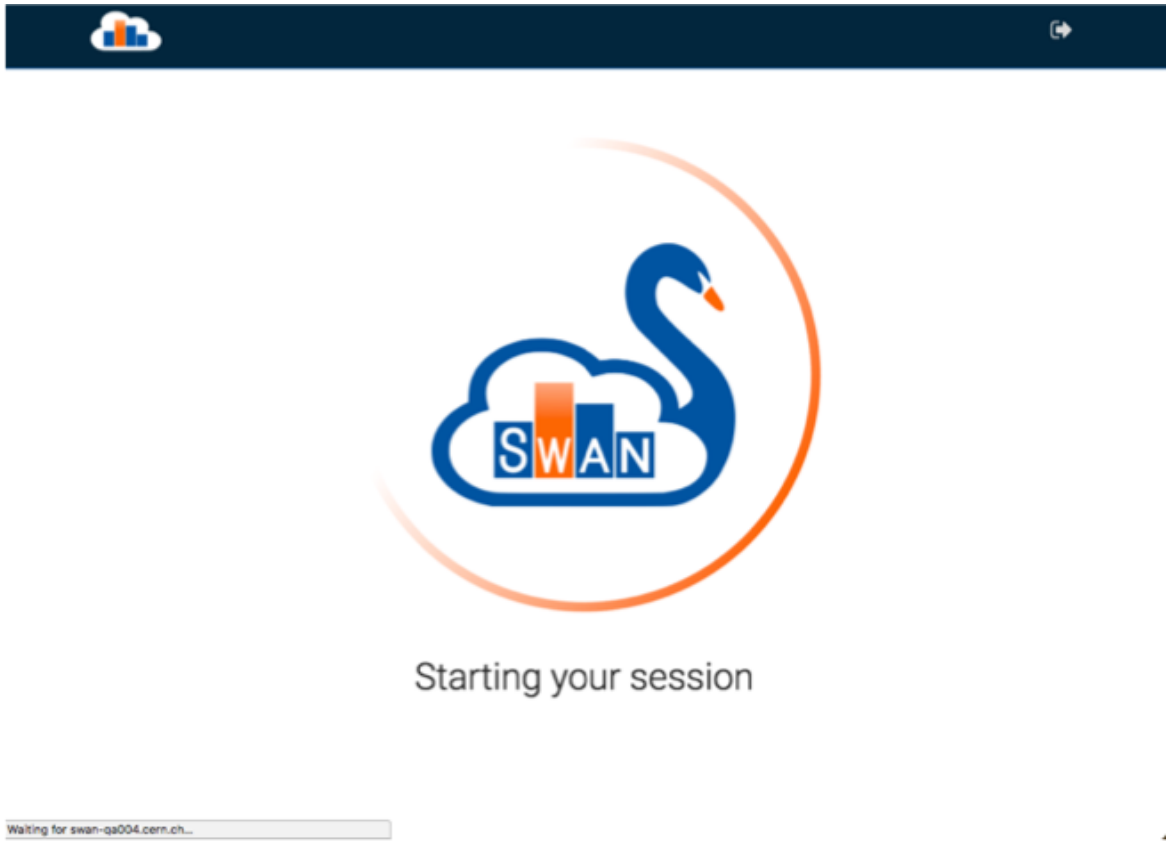
Boxed

- Containerized version of all the infrastructure
  - Includes EOS, CERNBox, CVMFS and all Swan services (Jupyter Docker image, JupyterHub)
  - Available in https://github.com/cernbox/uboxed
- Easily deployable on premises
  - Installable in Linux systems
  - Based on Docker Compose

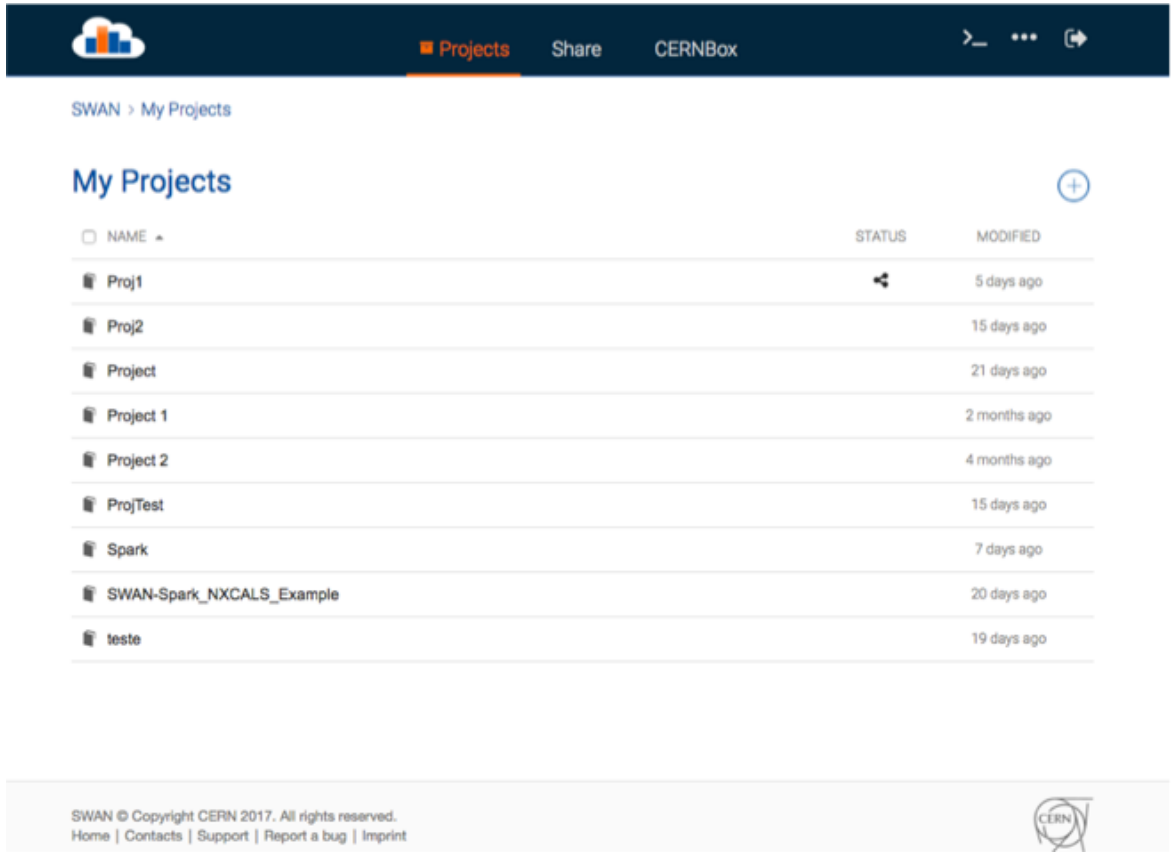
Recent developments

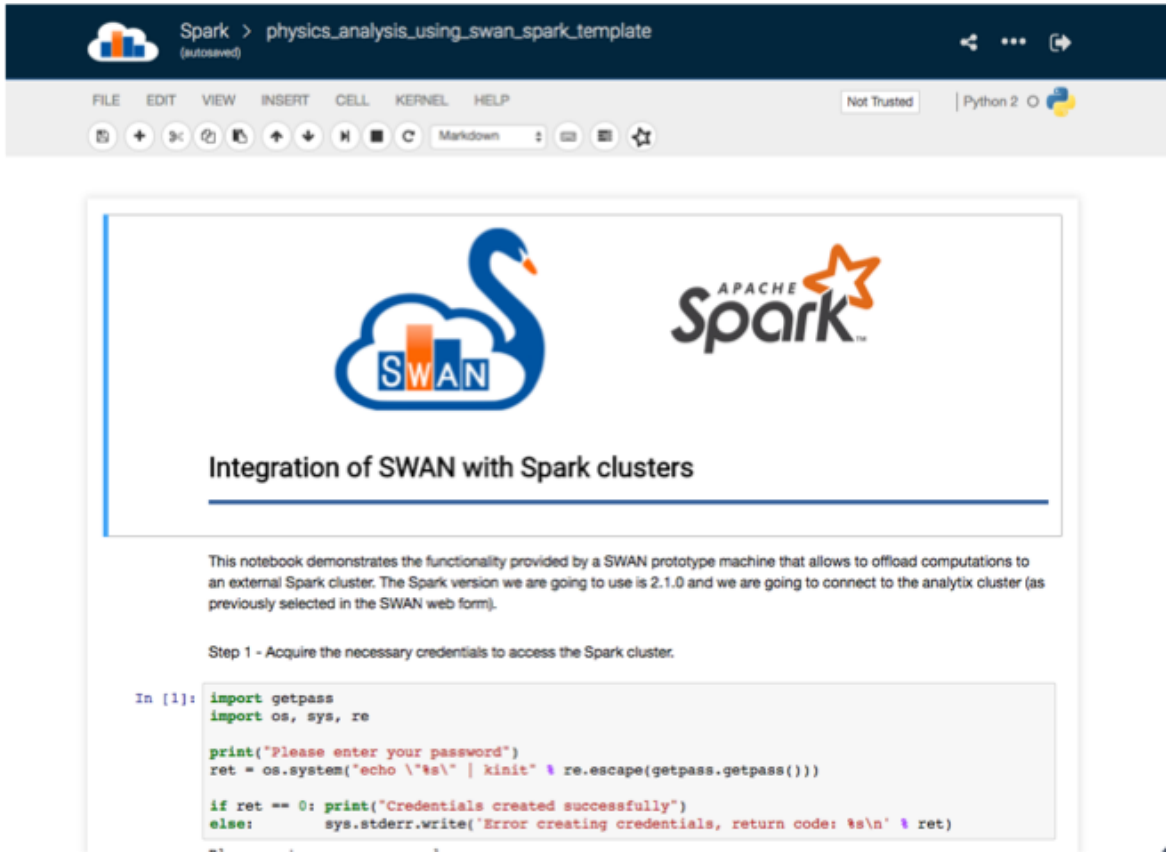
New User Interface





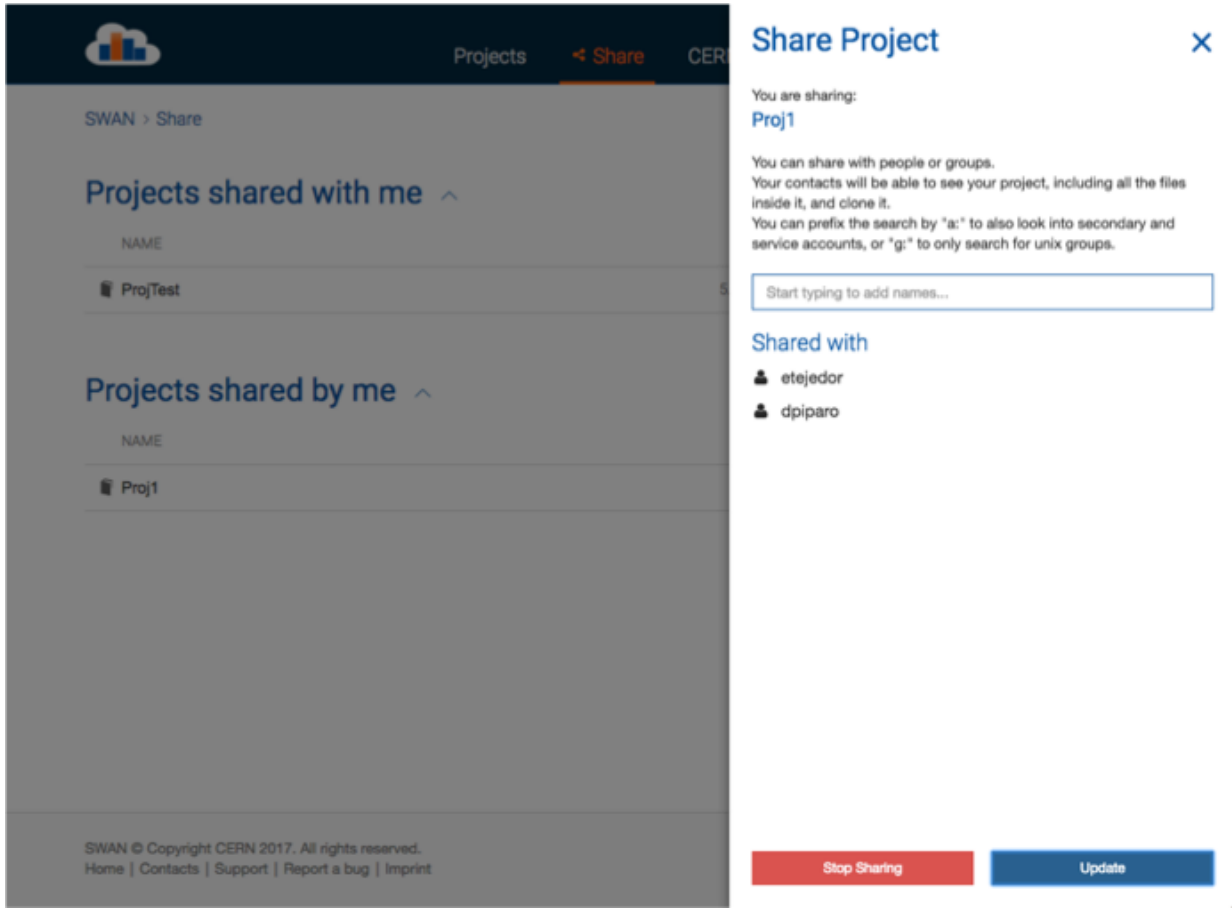
New User Interface





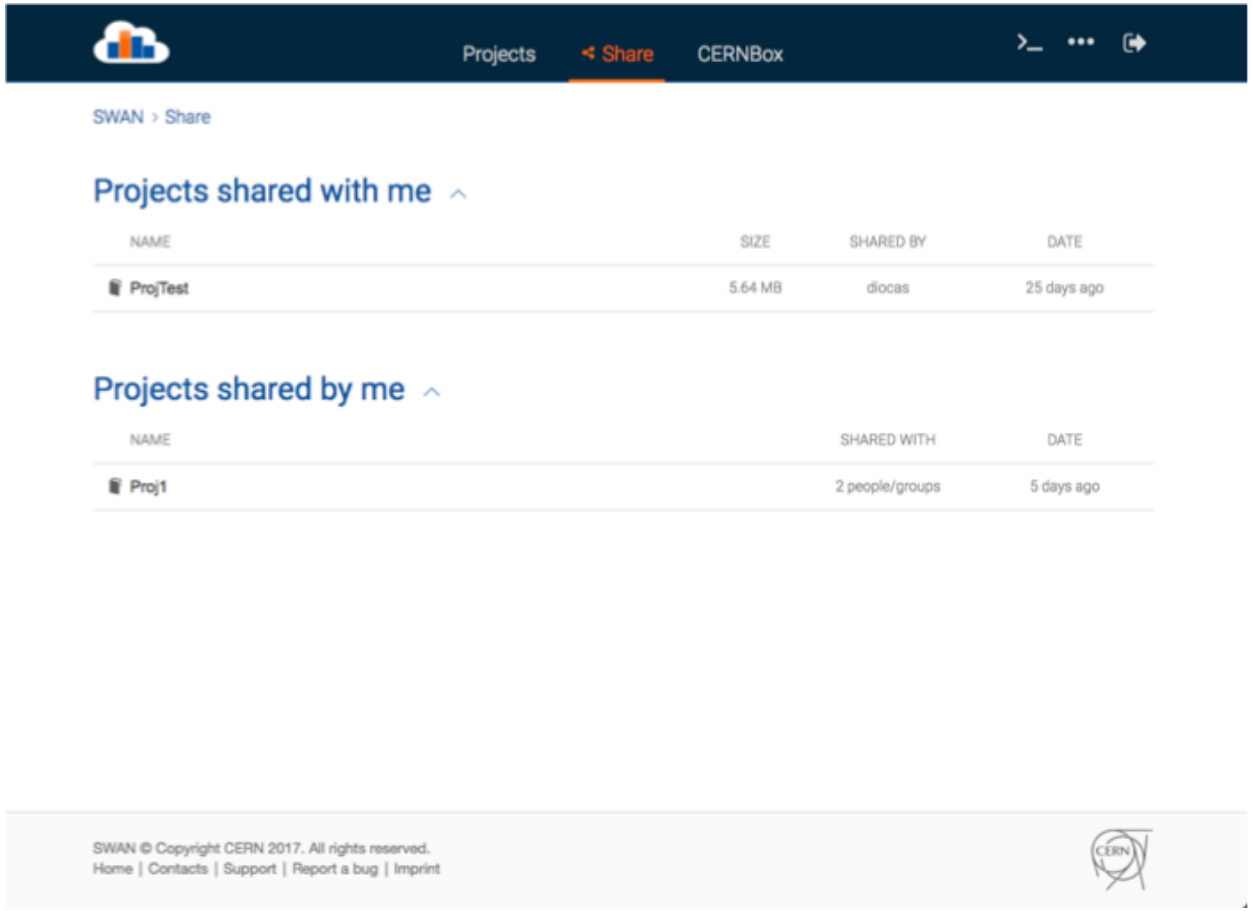
Sharing made easy

- Sharing from inside SWAN interface
  - Integration with CERNBox API
- Users can share “Projects”
  - Special kind of folder that contains notebooks and other files, like input data



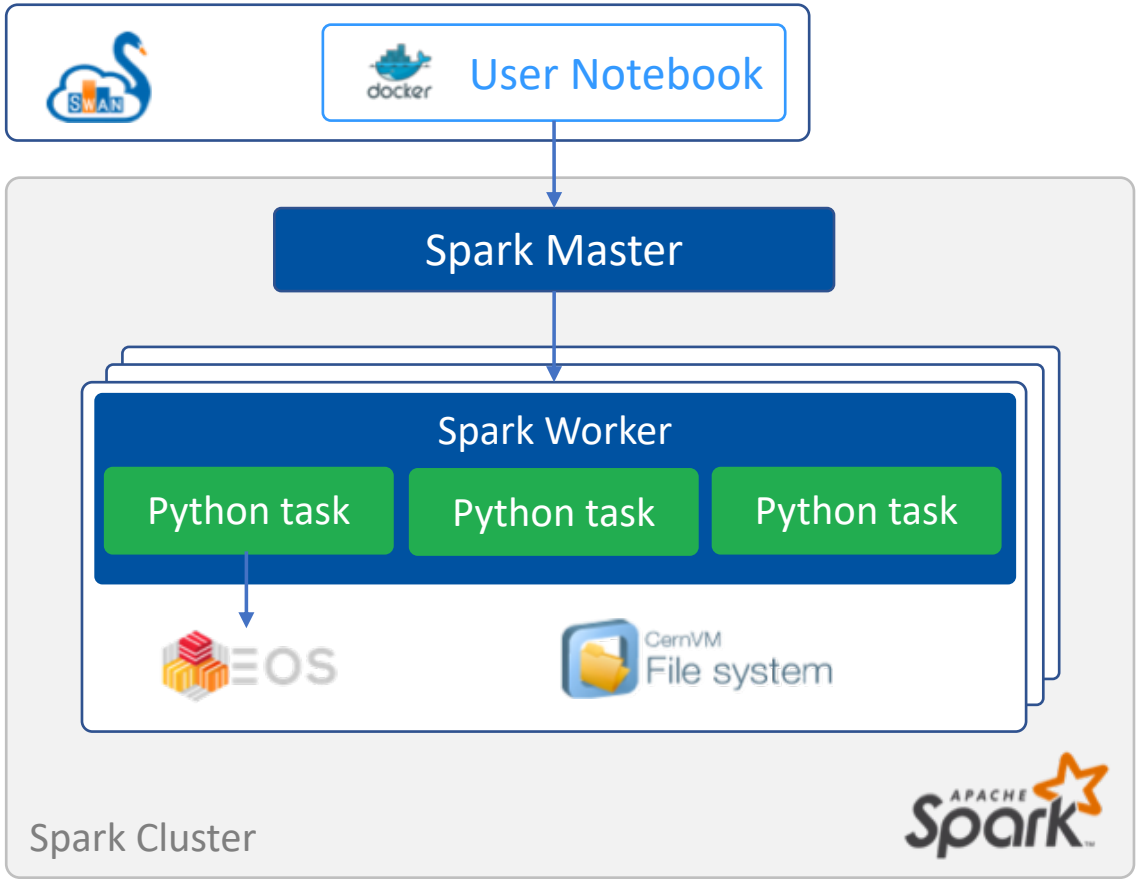
Sharing made easy

- Users can clone a shared Project directly from the interface
  - Jupyter doesn't allow concurrent editing

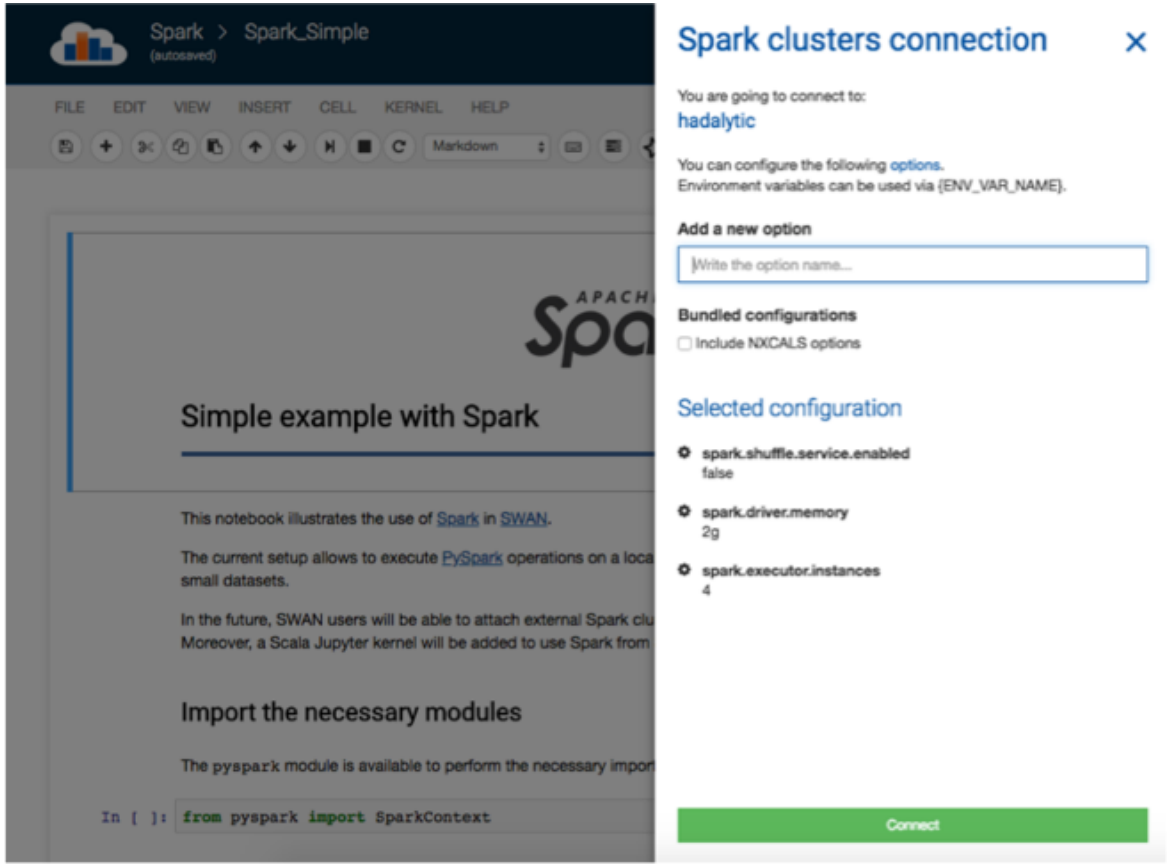


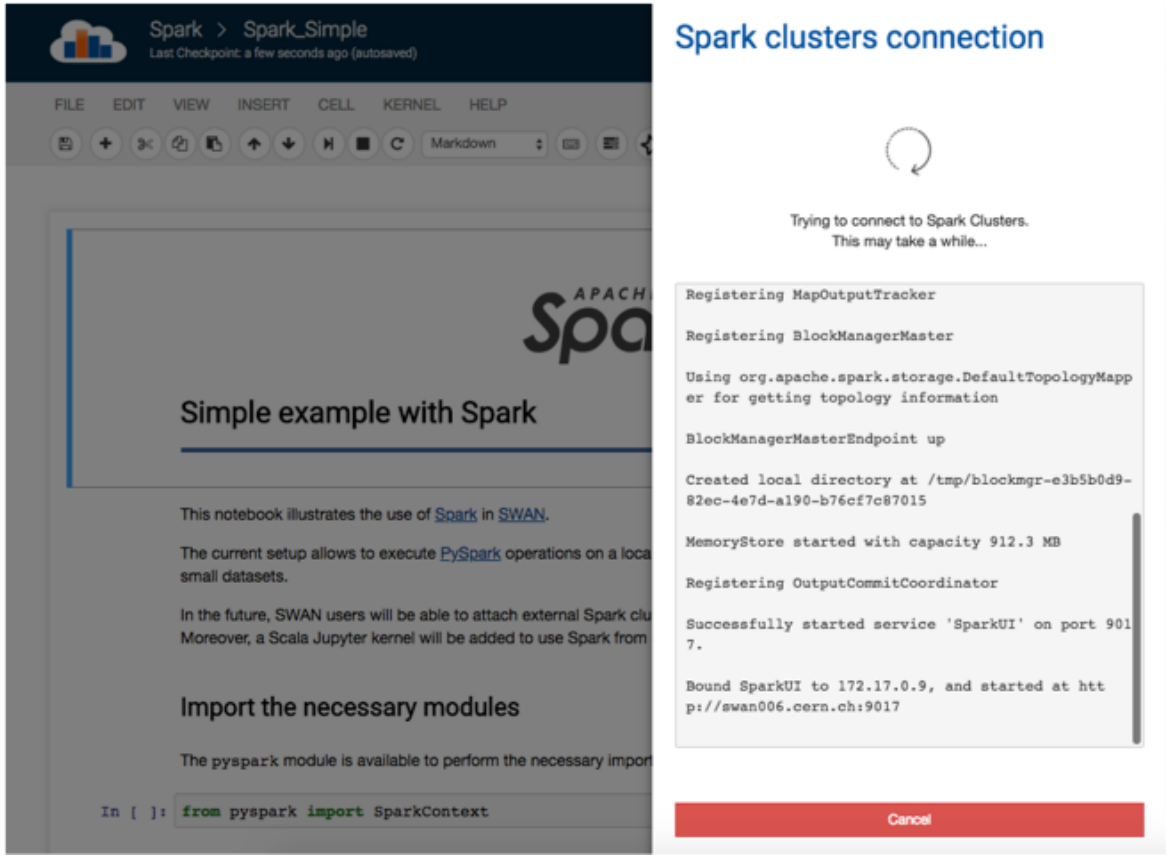
Integration with Spark

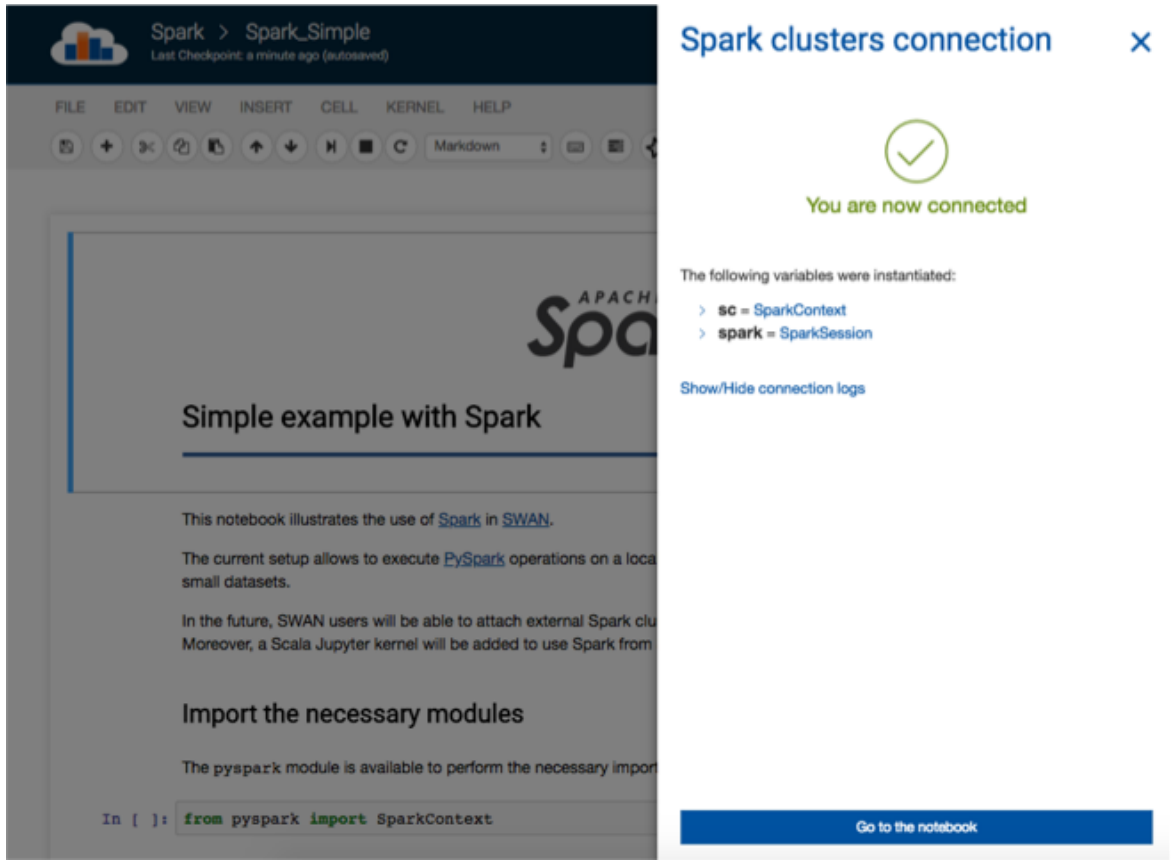
- Connection to CERN Spark Clusters
- User data accessed through EOS
- Graphical Jupyter extensions developed
  - Spark Connector
  - Spark Monitor



Integration with Spark



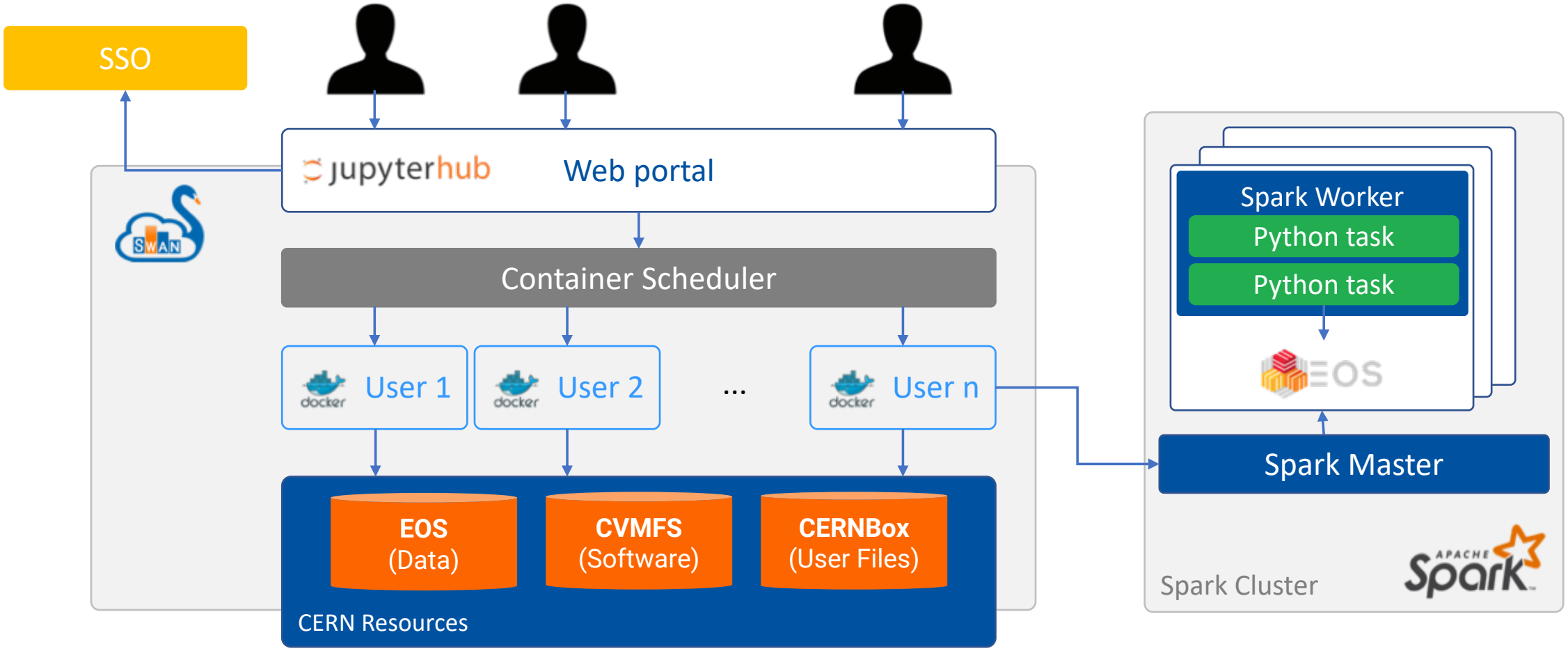



Spark > Spark_Simple
Last Checkpoint: a minute ago (autosaved)
FILE EDIT VIEW INSERT CELL KERNEL HELP
Markdown
Simple example with Spark
This notebook illustrates the use of Spark in SWAN.
The current setup allows to execute PySpark operations on a loca
small datasets.
In the future, SWAN users will be able to attach external Spark clu
Moreover, a Scala Jupyter kernel will be added to use Spark from
Import the necessary modules
The pyspark module is available to perform the necessary import
In [ ]: from pyspark import SparkContext

Spark clusters connection
You are now connected
The following variables were instantiated:
sc = SparkContext
spark = SparkSession
Show/Hide connection logs
Go to the notebook


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

Architecture

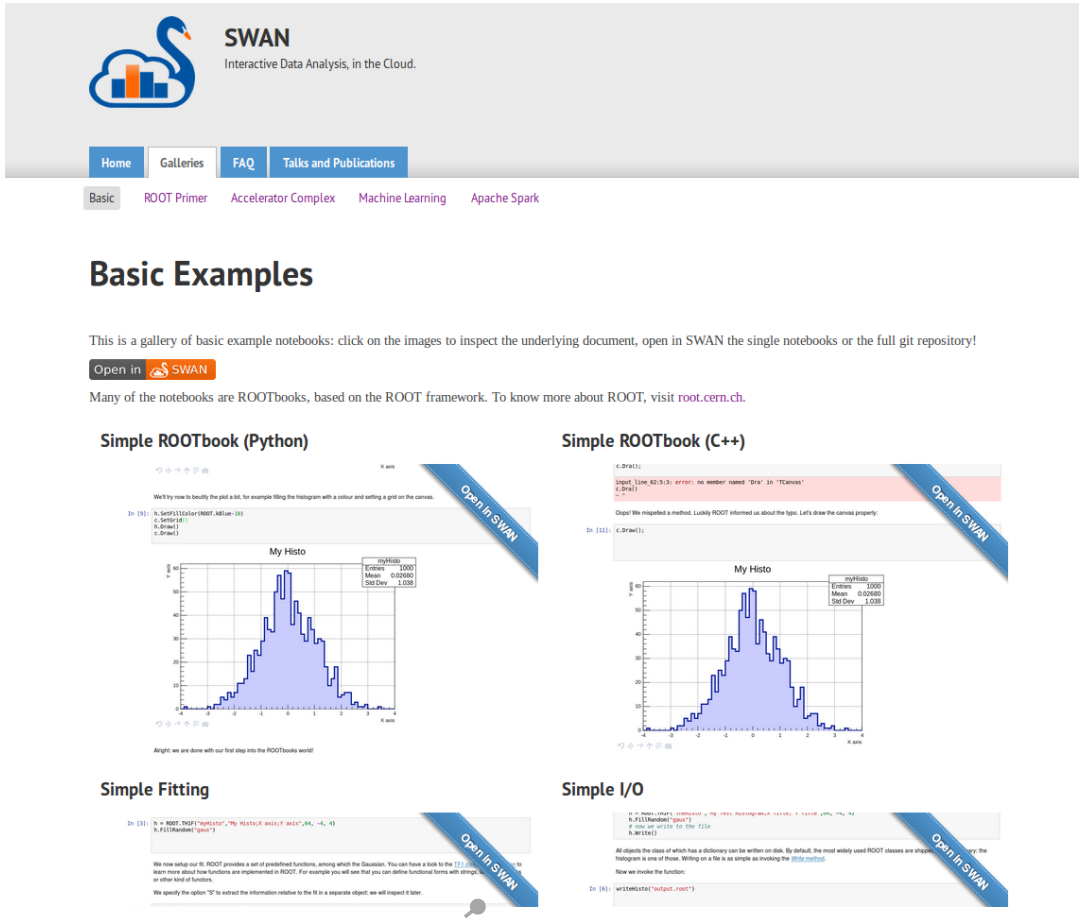


SWAN and the community

CERN 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
  - Almost 50 notebooks in 7 categories

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



Access with only a click

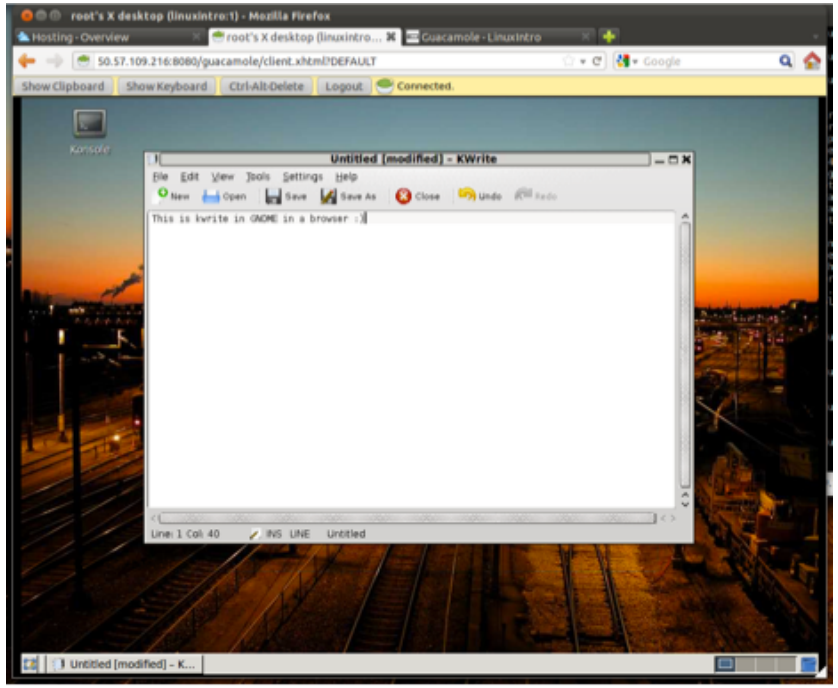
Other collaborations

- 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
  - Using Boxed distribution

Future research

Future research

- Apache Guacamole
  - Docker container available as remote desktop in a browser
  - Access to already configured software environment
  - More complex use cases not possible with Jupyter
    - i.e. access to ROOT framework
- Jupyterlab
  - More complex interface
  - Access to different editors
  - Concurrent editing



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