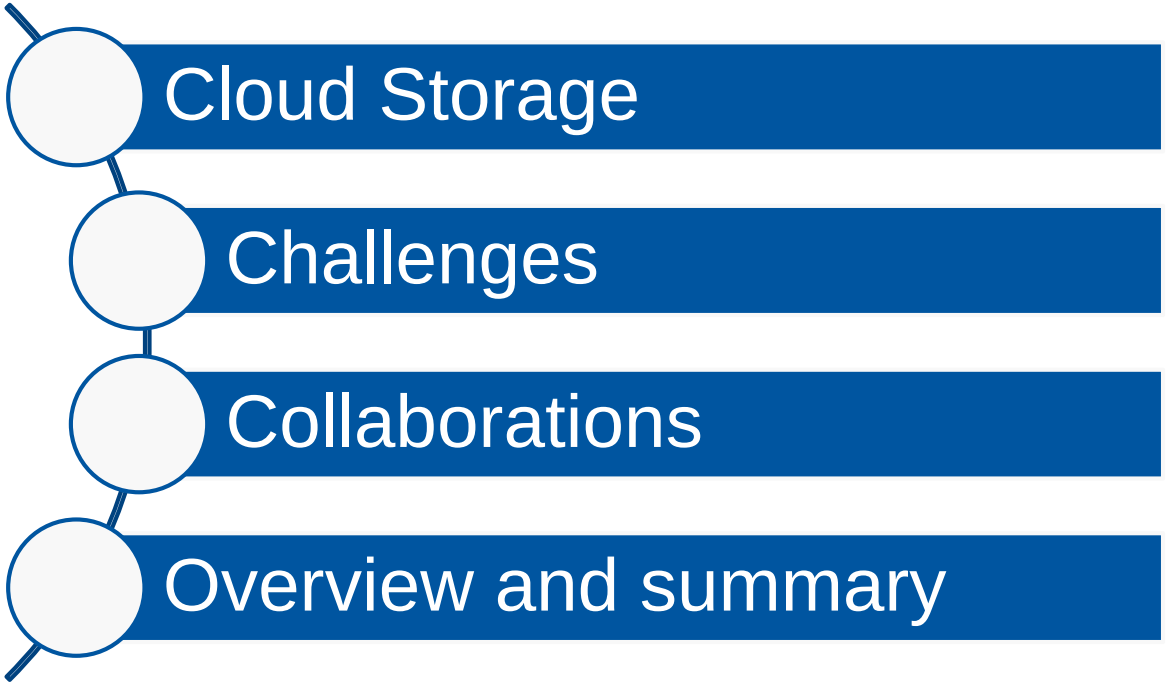
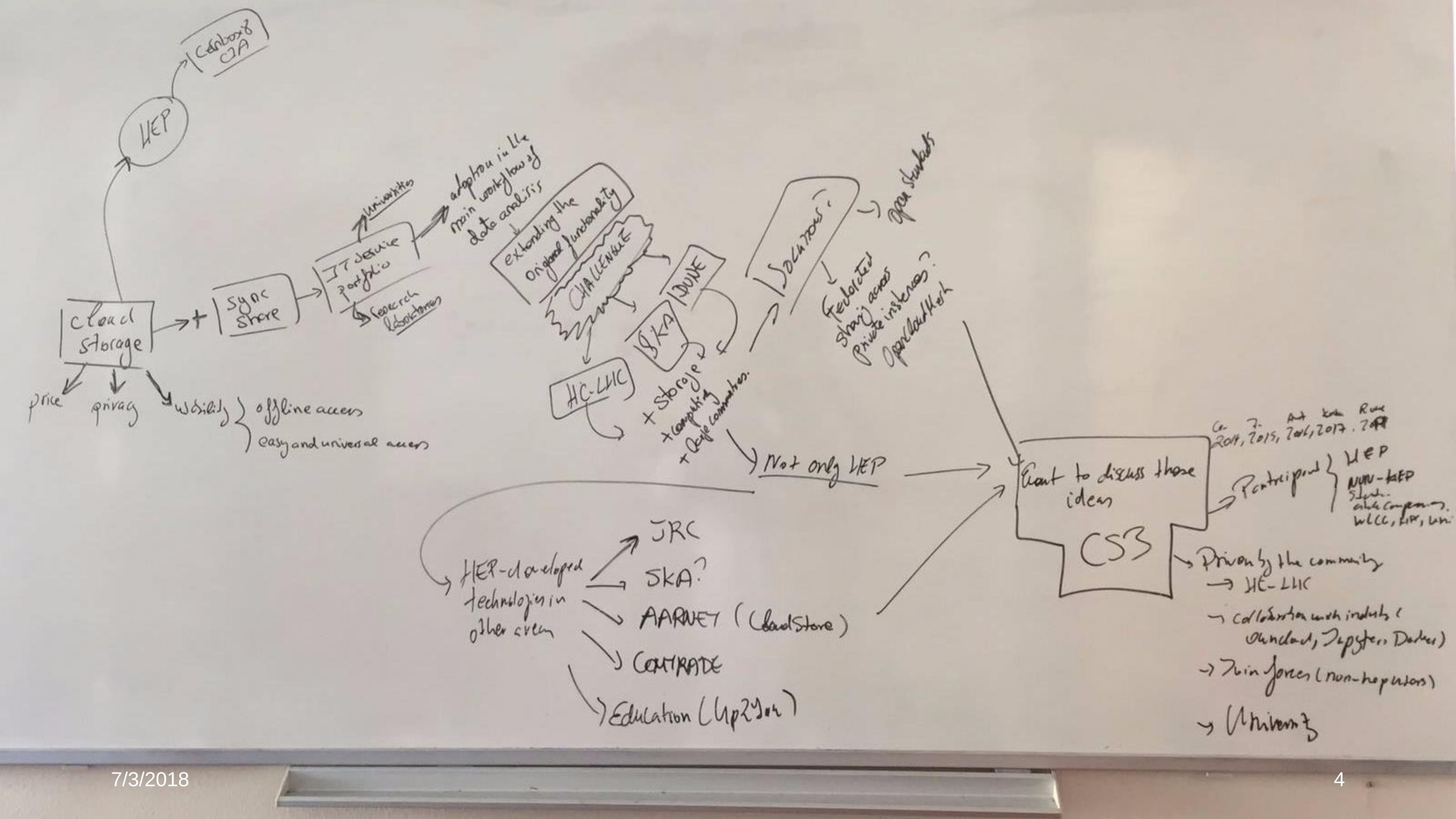

CERN

Cloud Storage for Data Intensive Sciences in Science and Industry

Hugo González Labrador – Storage Group

Outline


Cloud Storage
Challenges
Collaborations
Overview and summary


HEP
cloud storage
sync share
IT service portfolio
universities
research laboratories
adoption in the main workflow of data analisis
extending the original functionality
CHALLENGE
DUNE
SKA
HC-LHC
+ storage
+ computing
Solutions?
open standards
Not only HEP
price
privacy
offline access
easy and universal access
Event to discuss these ideas
CS3
2014, 2015, 2016, 2017, 2018
Participants
HEP
NON-HEP
Driven by the community
→ HC-LHC
→ collaboration with industry
→ Universities
HEP-developed technologies in other areas
JRC
SKA?
AARNET (CloudStore)
CERNBOX
Education

What do we see coming

- Interesting projects with new challenges
  - High-Lumi LHC, Dune, ...
- Just bigger?
  - e.g. HL-LHC x10 more luminosity/storage x50 more compute (more complex events)
  - New experiments: e.g. Dune
- New technologies!
  - Rethink the way HEP does data analysis
  - Bigger batch capabilities?
    - Natural sharing capabilities and seamless integration of large facilities with private resources

New technologies (1)

- Role "industrial" solutions
  - Marketed with different names
    - Cloud, Big-data, ...
  - Areas to get elements for our solutions
  - Areas for collaborations
- Second-level effect
  - User-base expectations
    - Why this is not as easy as {Dropbox, Facebook, } ?
    - Attracting brilliant students
  - More batch?
    - We definitely know how to build a larger computer centres
    - We should optimise the human part of the process (to make it more efficient):
      - Private copies vs sharing
      - Reinvent vs improve via reuse

New technologies (2)

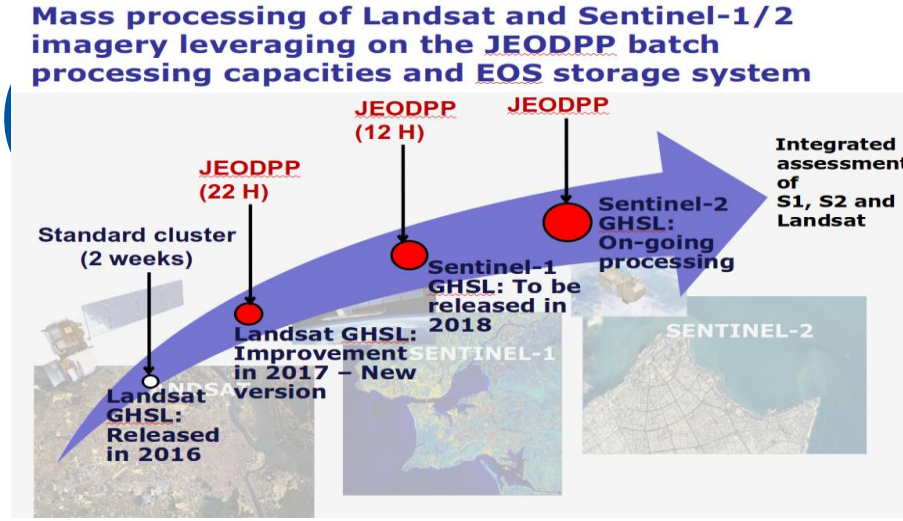
- Other sciences?
  - Relatively straightforward
    - We speak the "same" language
  - They are going through similar (re)volutions in the computing models
    - Data explosion
    - Compute explosion
    - Large distributed communities
- Examples
  - Earth Observations (e.g. EU JRC)
  - Astronomy (SKA -Square Kilometer Array- Australia and South Africa)
  - Support scientists proving infrastructure(CLoudstor - AARNET)
  - Reach out future scientists (UP2U)

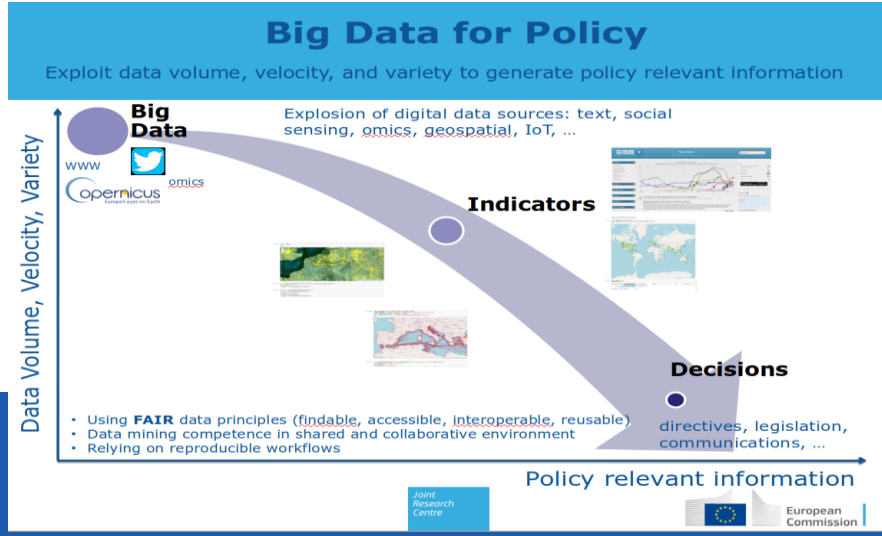
1 or 2 slides on

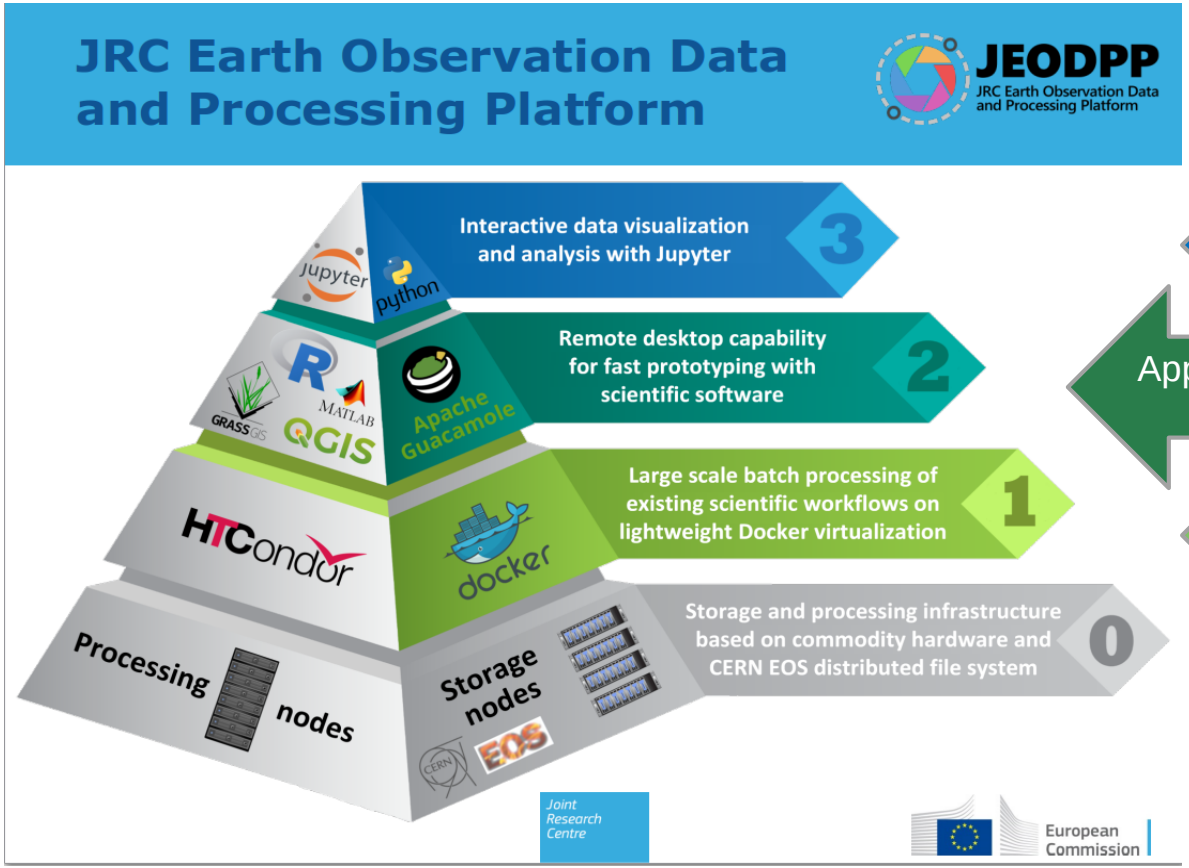
- 1 slide (just pictures and references to other CHEP talks)
  - EOS
  - CERNBox
  - SWAN (including the BE)
- 1 Slide
  - boxed on AWS
  - boxed on Helix
  - TOTEM

Earth observation (

- More data
  - More/better satellites
- Turnaround time
  - Importance of satellite data for everyday life
  - Data --> Actionable information
    - Agriculture
    - Pollution
    - Flood
    - Climate changes
    - ...
- Slides from C. Macmillan (JRC):
  - opening session at "Big data in Space",
    Oct 2017, Toulouse







Similar tools as SWAN

Application specific

Commonalities with WLCG sites (Docker,HTCondor)

Commonalities with WLCG sites (EOS, OpenStack)

P. Hasenohr and A. Burger,JRC presentation at CS3 2018 (Krakow)

Castille et al., FGCS, 2017, DOI: 10.1016/j.future.2017.11.007

CLOUDSTOR FILE SENDER + STORAGE

Collaborating nationally and internationally has got so much easier for AARNet customers. We've merged our two popular web services CloudStor (FileSender) and Cloudstor+ (cloud storage) into one easy-to-use solution.

CloudStor removes the frustration of slow data transfer rates and very large files by providing a super-fast, easy-to-use and secure file transfer and storage solution hosted on the AARNet network.

Unlike most cloud storage services, CloudStor is designed to meet the specific needs of researchers, and 100GB free storage is available to each individual researcher at AARNet-connected insitutions.

LOGIN TO CLOUDSTOR

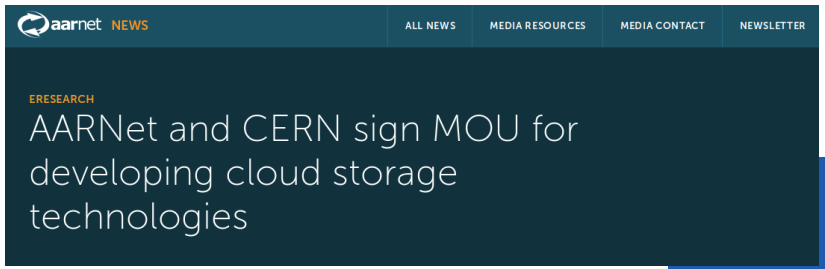
WHY CLOUDSTOR?

- 100GB free storage for individual researchers + group storage quotas for research projects.
- Quick and secure file transfer with no file size restrictions.
- Single sign on using home institution credentials (for Australian Access Federation members).
- CloudStor web interface for access to file storage, CloudStor FileSender and the AARNet Mirror.
- Storage located in Australia and directly connected to the AARNet backbone for rapid and convenient access, and avoiding any sovereignty issues.
- Data is replicated a minimum of three times at geographically distributed storage nodes for high reliability and availability.
- Cloudstor uses EOS, the scalable back-end storage developed at CERN.
- Sync client is available for Windows, Mac, OSX, Linux, iOS and Android.
- Access Amazon and other cloud data stores remotely using WebDAV and S3.
- Upload data sets from scientific instruments with CloudStor Rocket upload tool.
- Works with institutional repositories and national merit-based storage.
- A sustainable service that AARNet plans to provide indefinitely.

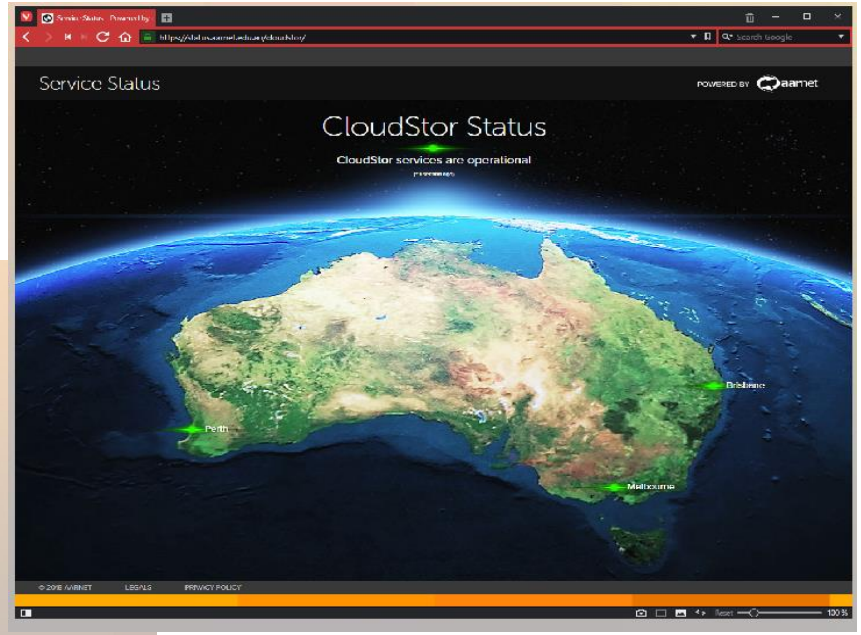
Similarities with some projects in HEP

Cloudstor (AARNET) is:

- **Ditributed from the start**
- **Multi-science from the start**
- **Multi-platform from the start**

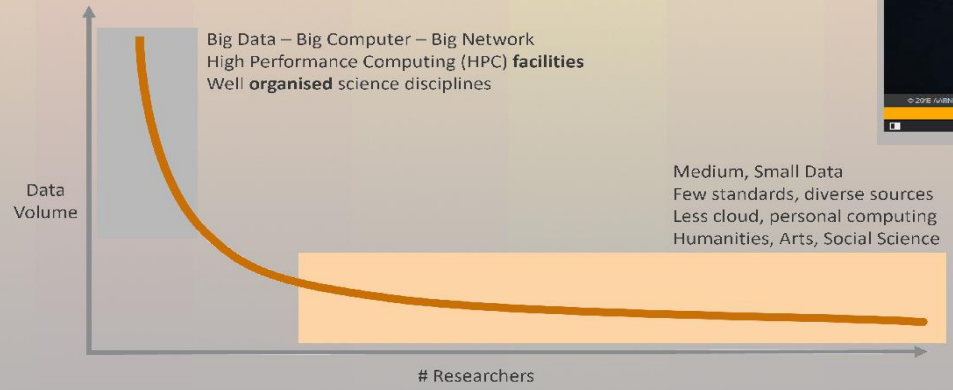


Cloudstor



Problem Scope

- Data sets researchers want to store are very different
  - Ephemeral data to archival data
  - Many small files to fewer very large files



Steal slide to Luca (Comtrade)

- Mention OpenLab collaboration

Steal slide to Kuba (UP2U)

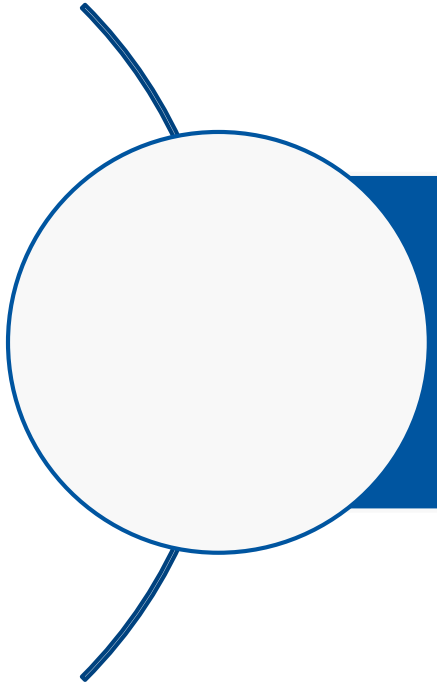
Steal slides to Kuba (CS3 publicity)

- History of companies at CS3
- Why CS3 is important (see next ones with some of my ideas, to be completed by you, notably by Kuba)

- "See you in Rome" slide

CS3

- Started as a workshop to learn (from each other, including academia and companies) how to provide cloud storage for scientific communities
- Right participant base
  - HEP and non-HEP
  - WLCG sites, HPC sites, University sites
  - Academics, start-ups, established companies
- We believe the drive of our community is an important factor of progress
  - Which is needed (cfr HL LHC)
  - Which can achieved by taking on board interesting technologies developed outside (cfr OwnCLoud, Jupyter, Docker)
  - Need to join forces (cfr collbration on EOS with non-HEP initiative as JRC and AARNET
  - University (UP2U) important as outreach but also as source of input (expectation of usage of our tools)



Overview and Summary

Conclusions

- HEP-developed software (CERNBox, EOS, SWAN) can boost the use cases of other sciences (JRC, AARNet) and constitute the backend for a new generation of services.

- The challenge is to adopt cloud storage solutions into the main data analysis workflow.

- Focus on human efficiency rather than only on machine performance, changes in the way people perform their analysis.