CMS considerations

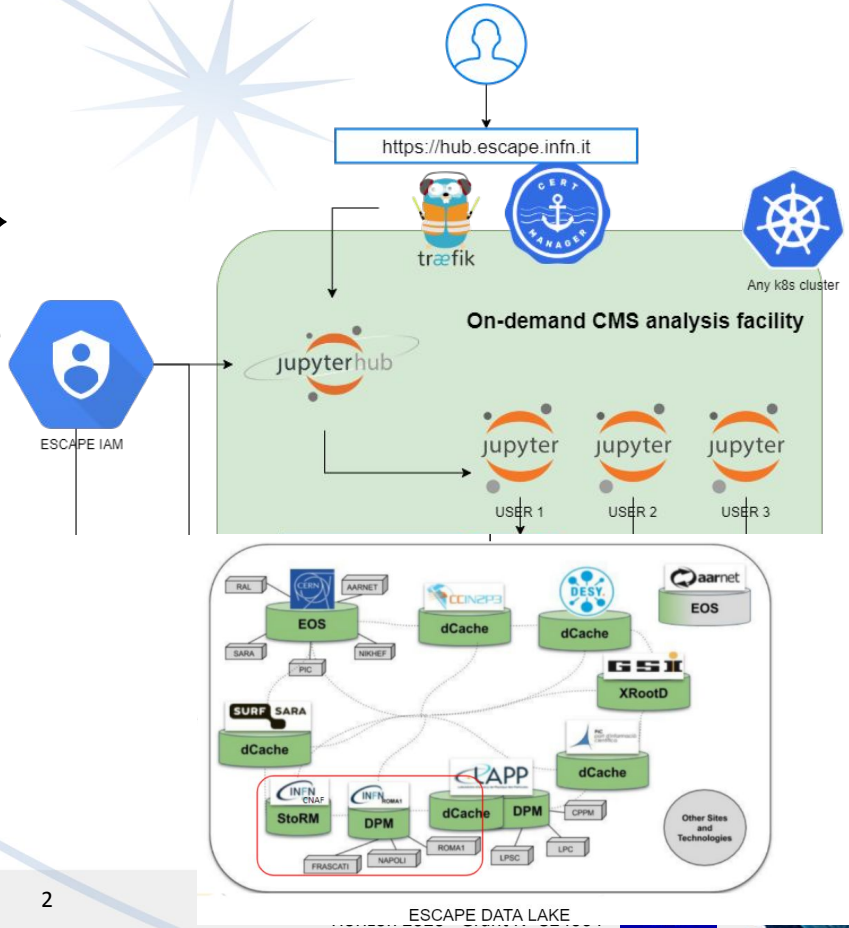
D. Ciangottini on behalf of the ESCAPE-CMS team

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

DAC21: recap

Brief recap on the main objectives for CMS:

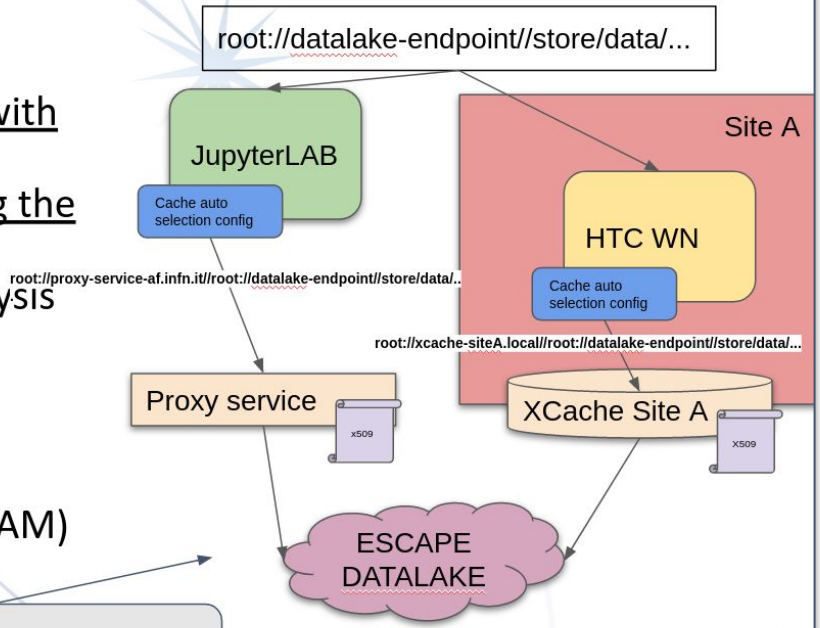
- **NanoAOD opendata**→ **plain ROOT files**→ python **ROOTDataFrame** as framework
- We were interested in replicating a simple analysis on a JupyterLab instance on a dedicated JupyterHUB hosted at CNAF
  - Also scaling out to a batch system
  - Demonstrate **scale out over HPC CINECA**
- We targeted opendata, but, of course, interested also to take a look at embargoed data management



DAC21 highlights 1/3

Access data via a group-managed Cache

- The cache server authenticates with a “service” proxy with the lake
- It serves the client either trusting the ip or with experiment IAM
- In other words, we perform analysis tests for which the user could in principle forget about any configuration to access the data (once is authenticated via a JupyterHUB and an experiment IAM)



SETUP DEPLOYED AT CNAF
and used for the following tests

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Funded by the European Union’s Horizon 2020 - Grant N° 824064

DAC21 highlights 2/3

Embargoed data

We also managed (credits to A.Ceccanti and L. Morganti) to give a first try to the upload of EMBARGOED CMS data to the CNAF-STORM RSE:

- Basically we succeeded to give **exclusive access** to the pfn reserved to the scope **CMS_EMBARGOED_DATA to the escape/cms users**
- The authorization is based on the user group both in the x509 and token!

In other words, all we needed to do was getting a valid x509 proxy or jwt with the voms/wlcg.groups attributes and then something like:

```
rucio upload --scope CMS_EMBARGOED_DATA --rse CNAF-STORM
--lifetime 90000 --summary --name ZZTo4mu.root ZZTo4mu.root
```

Configuration for other interested sites can be replicated, feel free to reach out!

Funded by the European Union's Horizon 2020 - Grant N° 824064

Funded by the European Union's Horizon 2020 - Grant N° 824064

DAC21 highlights 2/3

European Union's Horizon 2020 - Grant N° 824064

Things that worked and have been identified to help ESFRI/RIs on Data Management and Data Access

- Embargoed data can be enabling for the adoption in scenarios where access to data might be restricted to members of a group
- RUCIO authN integration with OIDC was a key for the integration in a modern analysis infrastructure
  - We will keep pushing in further improvements in this area
- RUCIO Jupyterlab plugin

Identified barriers to adopt the DIOS model, services or tools

N.T.R. also because for most of the problems we found a very good support and great community participation

Funded by the European Union's Horizon 2020 - Grant N° 824064

Your ESFRI/RI plans regarding the technologies exposed in DIOS, will you consider continue exploring or adopting?

As CMS we would like to look at:

- How can RUCIO manage replication of “embargoed” data on a more heterogeneous set of storage technologies (different technologies)
- Getting some preliminary experience with ephemeral/volatile RUCIO RSEs for buffer/staging areas
  - user analysis cached inputs
  - Dynamic output registration

Did you identified security issues? Any specific security related worries to name (present or future) ?

We are interest in extending the security model for secret/token management of our JupyterHub instance with Hashicorp Vault (or equivalent solutions).

Suggestions for general improvements on DIOS: model, services, tools, etc.

N.T.R.

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Is your ESFRI/RI Interested in a longer term existence of an ESCAPE or an ESCAPE-*like* infrastructure?

As CMS we look forward to the implementation of a data-lake model in line with the one tested with ESCAPE project. In fact the studies on the future models for data access are already started, although the final adoption is obviously led by WLCG decisions at large.

Also, at national (INFN) level, some initiative are ongoing and they are committed to the evaluation of a datalake model over italian sites.

In this sense interoperability is, in our view, a key point to share among all the future initiative

Is your ESFRI/RI interested establishing standing collaborations, channels, joint efforts? On which specific topics?

As CMS we adopted RUCIO as the core component of the DataManagement infrastructure, hence collaboration in this respect can/will follow the official RUCIO channels

Even more interesting would be to focus on establishing a joint effort to design common solutions that would facilitate the interoperability of a scientific data lake.

Funded by the European Union's Horizon 2020 - Grant N° 824064