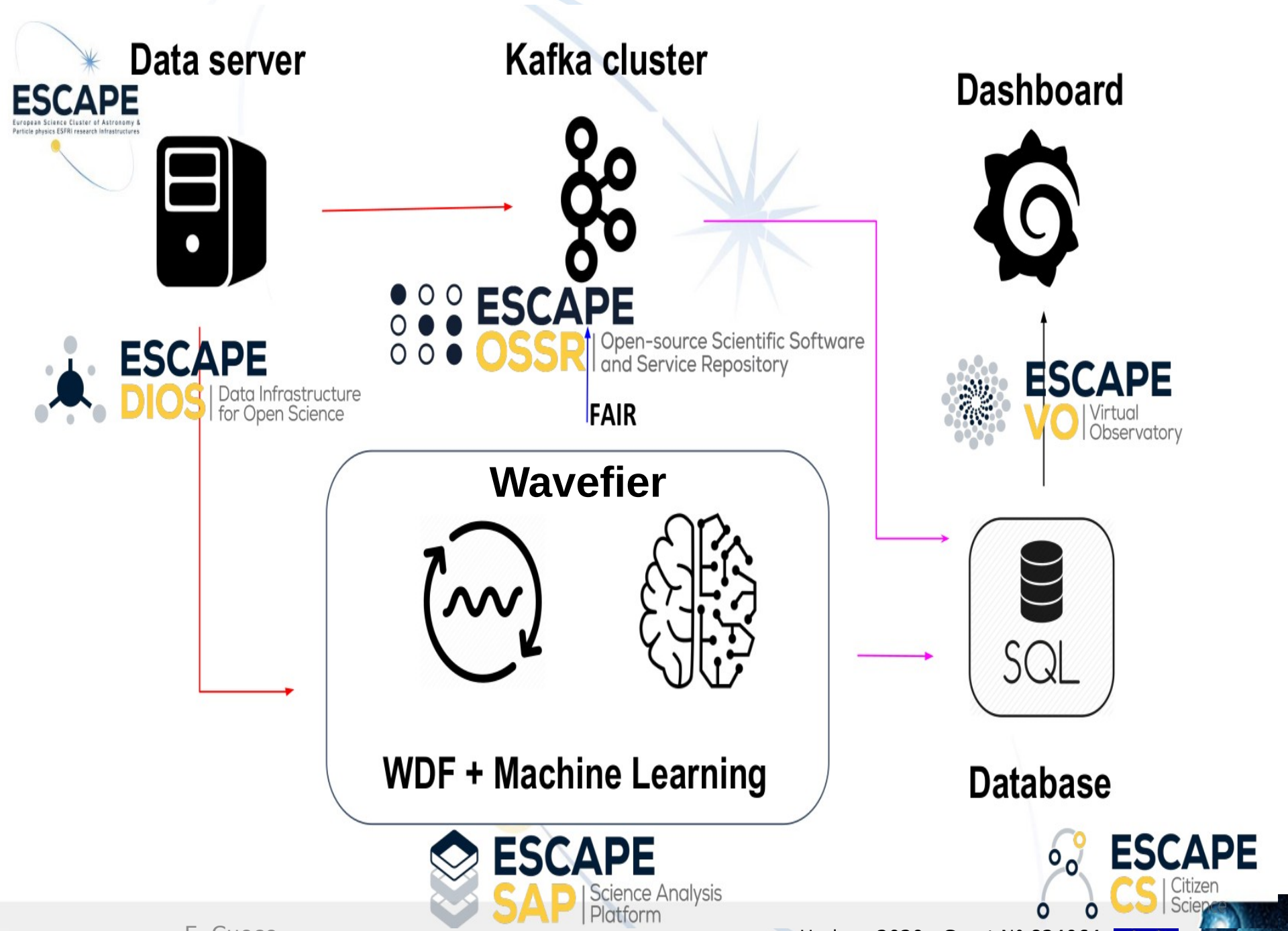
Dress Rehearsal Review
EGO / Virgo perspective

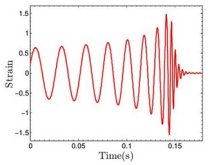
Pierre Chanial, Sara Vallero, Elena Cuoco, Filip Morawski

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



Horizon 2020 - Grant N° 824064

Wavefier Online / Offline Architecture



EGO server
Streaming public data
in real-time

ESCAPE data-lake

Other potential
real-time or offline
pipelines

CNAF Kubernetes Cluster

Datalake
Importer

kafka
timelines
triggers

Wavefier
WDF
(detector)

Wavefier
ML
(classifier)

Grafana
dashboard

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

First Data-lake Injection Dress Rehearsal

- EGO : 4h test of upload to and download from the ESCAPE data-lake prototype
  - real-time
  - public Virgo h(t) drawn randomly from an O2 h5f file
    - O2 llhoft have not been made public
    - ESCAPE data-lake prototype not yet secured for proprietary data
  - chunks : 1 second, 4kHz
  - Data rate : 85kB / s

- Goal is to test functionalities, not yet performances. Latencies have been measured, but they should be taken as the baseline we will improve on, not definitive numbers.

- Uploader: Celery application to pace the uploads
  + Rucio Python non-docker client

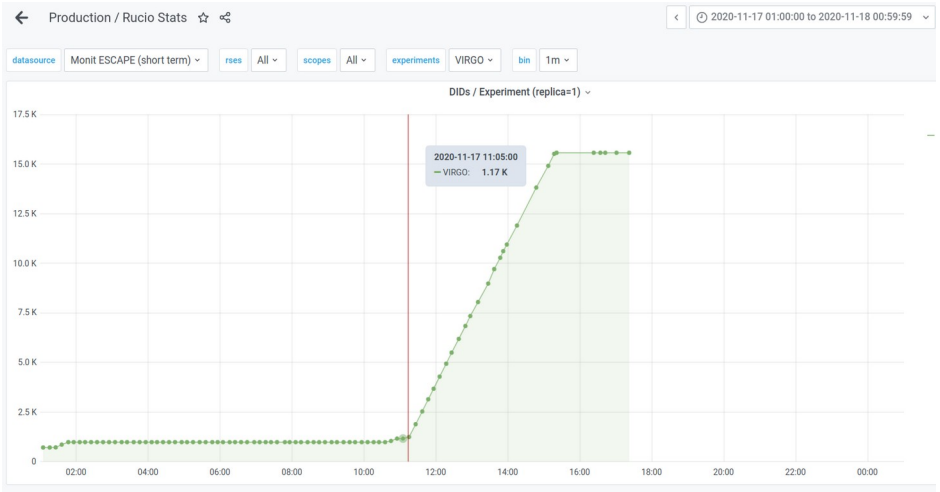
- Downloader: Multiprocessing Python
  + Rucio Python docker client

  Layman's approach : download the dataset content metadata at regular intervals to poll new entries in the dataset

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

First Data-lake Injection Dress Rehearsal

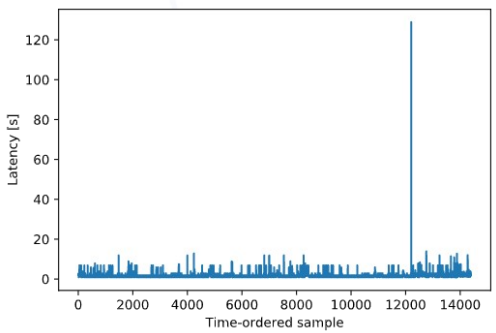
- Great success
- The 4 x 3600 = 14400 data chunks have been sent
  - ➔ All samples uploaded
  - ➔ All samples downloaded
  - ➔ None corrupted

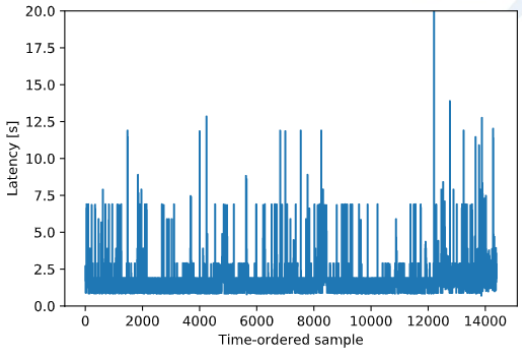


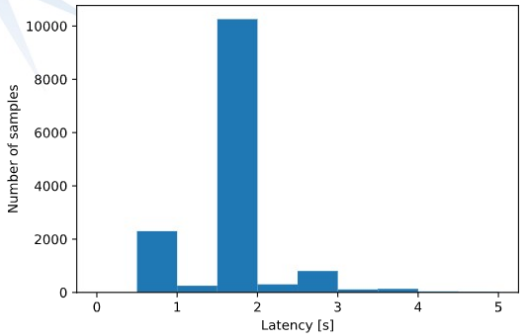
Funded by the European Union's Horizon 2020 - Grant N° 824064

First Data-lake Injection Dress Rehearsal

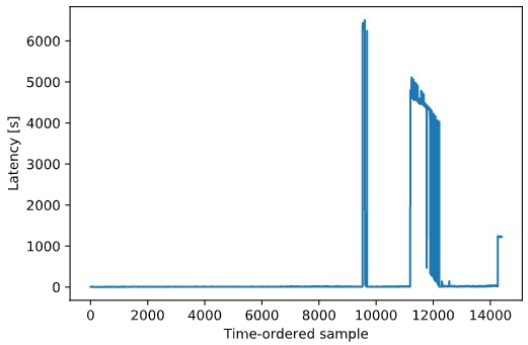
- Upload (EGO ➡ CERN) latency analysis (mean : $1.9s_{-0.5}^{+0.3}$ s)

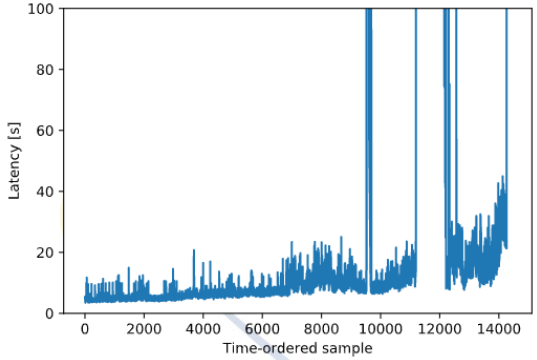


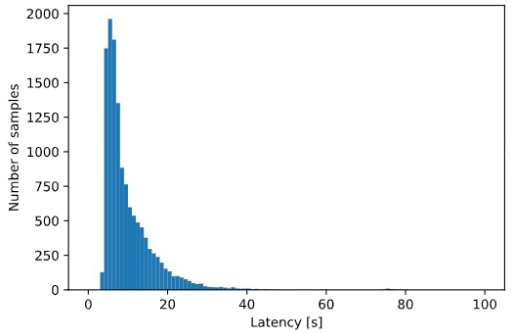




- Total (EGO ➡ CERN ➡ CNAF) latency analysis in progress (median : 8.2 s)
  Downloader can be vastly improved.







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

Second Data-lake Injection Dress Rehearsal

- 15th of December
- Hopefully : use of Rucio events for notifications of new files

  but not for the general public, needs access to Rucio's ActiveMQ broker
- Otherwise : another scenario : replication of Virgo raw data in two data centers for archival purposes
- Replication rules in two data centers
- Improved timestamp resolution

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

Data-lake Injection Dress Rehearsal Feedback

- What are the current and future data management concerns for your experiment? For future concerns, please give an estimate of timescales.

  Show-stopper: protection of proprietary data. Even if EGO/Virgo and LIGO are currently not acquiring science data until at least June 2022, a stream of replayed data is currently made available to real-time pipeline developers, and this stream is also proprietary. For archival purposes, we would require that the data be kept confidential, until it becomes public.

- What features of the WP2 datalake are most interesting from the perspective of your experiment?

  Automatic replication in data centers, QoS, Easy to upload / access / download files

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

Data-lake Injection Dress Rehearsal Feedback

- What is the scale of the challenge your experiment will face? Please include approximate annual data volume, number of files, number of sites, number of users.

  raw data : 1.6PB in ~500000 files
  strain data : 2.7 TB in ~30e6 files
  2 (currently) to 4 data centers
  order of 30 users.

- Based on your experiences so far, can you see how aspects of the datalake technology stack could be useful, and can you identify areas that need to be improved? ((i) Rucio, (ii) IAM, (iii) Monitoring, (iv) QoS, (v) Other)

  Rucio : easily accessing events (filtered by experiment by the server) to be notified of new files available in a dataset.

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