Coffea-casa news and development

Oksana Shadura
On behalf of coffea-casa development team

https://indico.cern.ch/event/1098937/
Coffea-casa Analysis Facility


Grid / cluster site resources
Kubernetes resources
Data delivery services - ServiceX
HTCondor scheduler
HTCondor workers
JupyterHub (shared between users)
Jupyter kernel
Dask workers
Dask scheduler
XCache
Skyhook
Remote data access
Data delivery services - ServiceX

Coffea-casa Analysis Facility


Grid / cluster site resources
Kubernetes resources
Per-user resources
Shared resources between users
Coffea-casa deployment: existing coffea-casa AF

- Coffea-casa style AF facilities, goal of adding more sites as soon as we gain experience

CMSAF @T2 Nebraska
  “Coffea-casa”
  https://coffea.casa

OpenData AF @T2 Nebraska
  “Coffea-casa”
  https://coffea-opendata.casa

ATLAS AF @Scalable System Lab (UChicago)
  “Coffea-casa”

New facility with ATLAS IAM, setting this up generated valuable feedback for future coffea-casa developments.
Analysis Grand Challenge

**Motivation:**
- Allow coping with HL-LHC data sizes by rethinking data pipeline
- Provide flexible, easy-to-use, low latency analysis facilities

Analysis Grand Challenge will be conducted during 2021–2023, leaving enough time for tuning software tools and services developed as a part of the IRIS-HEP ecosystem before the start-up of the HL-LHC and organized together with the US LHC Operations programs, the LHC experiments and other partners.

_Coffea-casa vCHEP 2021 plenary talk_
The idea behind the Analysis Grand Challenge

Analysis Tools

- **uproot**
  Reading and writing ROOT files (just I/O)
- **Awkward Array**
  Manipulating arrays with nested structure (not HEP-specific)
- **hep-tables**
  DataFrame for nested structure
- **Particle**
  2D, 3D, & Lorentz vectors
- **vector**
  Histogramming
- **uproot**
  Histogramming
- **iminuit**
  Raw minimization
- **zfit**
  Curve fits
- **hepstats**
  Statistical tools
- **mpthep**
  Plotting
- **HistFactory-style fits**
  Statistical models

Analysis Facilities

- **Coffeea-casa Interactive Analysis Facility**
- **JupyterHub**
  (shared between users)
- **Dask scheduler**
- **Jupyter kernel**
- **Dask workers**
- **HTCondor scheduler**
- **HTCondor workers**
- **Data delivery services - ServiceX**
- **Skyhook**
- **XCachex**
- **Remote data access**

Grid / cluster site resources
- Kubernetes resources
- Per-user resources
- Shared resources between users
IRIS-HEP AGC Tools 2021 Workshop, Nov 3–4th 2021

- Workshop showing **IRIS-HEP toolchain at coffea-casa instances**, aimed at PhD / postdoc level
  - [https://indico.cern.ch/e/agc-tools-workshop](https://indico.cern.ch/e/agc-tools-workshop)
  - 2 afternoons CERN time (15:30 - 19:30) on Nov 3/4
  - Brief introductions to individual packages, notebook talks focusing on interfaces between tools
  - Using **Open Data** examples, then splitting into ATLAS / CMS - specific tracks
    - Interest from **LHCb & smaller experiments**: encouraged follow-up meetings to understand their needs better

- **102 registered** participants
  - Closed registration because we were not sure if available AF resources would be able to host more participants

- **81 people connected** to Zoom on first day

- Event recorded & to be shared on Youtube

Materials:
Analysis pipeline demonstration

- Showed ecosystem integration with **analysis pipeline example**
- Interfaced many packages, deployed on **Open Data coffea-casa**
- Analysis example is small (~500 MB input, ~1 M events), but:
  - This approach scales! To be shown at future events.
  - Systematic uncertainties included — captures realistic features

**full notebook:** HZZ_analysis_pipeline.ipynb
Next milestones

- Dask dependency management plugin for coffea-casa
- Add CVMFS to be available from notebook
- Create mailing lists for OD and CMS AFs @ UNL
- Preparation for ACG Tools: Skyhook
- Deploy monitoring for coffea-casa @ UNL: prometheus/elasticsearch instances
- Provide unique HTCondor users for each coffea-casa user
- Add deployment documentation
- And many more!
Coffea-casa AF statistics

CMS instance:

- 27 active* users (some of them ~ 14 are using AF connected in the last 24 hours)
  - We had in total 100 users who ever used coffea-casa
- 8-9 git repositories with analysis tried on coffea-casa on a different stage of development (I found in Github)

* - total stats over last month

Please help us to become better:
[https://github.com/CoffeaTeam/coffea-casa/discussions](https://github.com/CoffeaTeam/coffea-casa/discussions)