

Analysis Grand Challenge workshop closing

Alexander Held (University of Wisconsin–Madison)

Oksana Shadura (University Nebraska–Lincoln)

May 3–5, 2023

IRIS-HEP AGC workshop 2023

<https://indico.cern.ch/e/agc-workshop-2023>



Workshop outline

A packed program over 2.5 days


- **Wednesday**
 - Morning: Analysis Systems & demos
 - Afternoon: focus on ServiceX
- **Thursday**
 - Morning: DOMA & facilities
 - Afternoon: facilities & future directions
- **Friday**
 - Morning: planning towards an AGC showcase event

AGC execution event

- We are working towards an “AGC execution” event
 - Likely to happen September 14 — stay tuned
- **AGC execution** will be a short, half-day event
 - Inviting everyone who is interested to **share setup and present results**
 - Interesting combinations of hardware, network site configurations
 - Any type of “combinatorics” of AGC implementation / components setup
 - Can include performance measurements
 - Chance to showcase your computing resources to physics analysis community :-)

AGC versions

Overview of current & future versions ([documentation](#))

- 
- **v0.1:** the **ACAT setup** ([related talk](#)), using ntuple inputs¹
 - version of current RDF implementation² (fellow project this year: move implementation to v1/v2)
 - **v0.2:** same analysis as v0.1, improved **ServiceX pipeline** (coffea streaming files from object store)
 - **v1.0:** switch to **NanoAOD** inputs — replaces all v0. Minimal analysis changes (new column names)
 - **today:** working towards v2: added **ML** training, MLFlow / Triton integration, **correctionlib** adoption
 - **v2.0:** (~ mid June target) **AGC for execution event. ML + more systematics** (increased I/O + CPU)
 - Aim to provide implementation with coffea 2023; if needed, move that to v2.1
 - **Execution event:** targets **v2.0**. Demonstrations based on v1.0 from other interested participants as backup.

¹ (ntuples_merged.json — no point in using the older ntuples.json)

² currently misses statistical inference part of pipeline

AGC pipeline configuration for execution event

What we would like to see in contributions

- **Baseline:** full AGC pipeline with **Dask** (`USE_DASK = True`)
 - Can also be ROOT version with distributed RDF
- **Advanced:** demonstrate pipeline with **ServiceX** (optional)
 - `USE_SERVICEX = True`,
 - Employ your XCache if available and compare performance
- **Advanced:** include **additional ML functionality** (optional)
 - Training: run `jetassignment_training` & reproduce models, more advanced: `USE_MLFLOW = TRUE`
 - Inference: `USE_TRITON = TRUE`

Options on this slide refer to the [ttbar_analysis_pipeline.ipynb](#) implementation.

Advanced performance studies

Additional aspects available for studies

- Execution event target for facilities: demonstrate baseline setup
- **Additional functionality** provided for more studies
 - Variations in I/O requirements for benchmarking (IO_FILE_PERCENT)
 - Turn on/off ML inference & columnar calculations (USE_INFERENCE, DISABLE_PROCESSING)

AGC execution event

Metrics that might be of interest

- **Goal** of execution event: **showcase functionality**, but welcome to use existing setups for more beyond that!
- **Standard metrics** (in the many configurations outlined previously)
 - Data volume processed (per time and core)
 - Event processing rate per core
 - Scheduling efficiency à la [David Koch's slides, page 12](#)
- **Data pipeline comparisons**: ratio of ServiceX+coffea and coffea (directly reading original input) runtimes
 - Assumption: input data sitting in XCache
 - Goals: no substantial slowdown of initial execution of ServiceX+coffea setup, demonstrate significant speedup in repeated runs (hitting ServiceX cache)
- **Additional points of interest**
 - Capture multi-user setups: run multiple AGC pipelines in parallel
 - Evaluate UX: how much manual intervention is needed (e.g. copying & settings tokens)

CHEP talks next week

AGC-related talks of interest

- David Koch: [Analysis Grand Challenge benchmarking tests on selected sites](#), Monday 12:15
- Elliott Kauffman: [Machine Learning for Columnar High Energy Physics Analysis](#), Monday 14:00
- Andrea Sciabà: [I/O performance studies of analysis workloads on production and dedicated resources at CERN](#), Monday 15:00
- Oksana Shadura: [Coffea-Casa: Building composable analysis facilities for the HL-LHC](#), Tuesday 10:00
- Alexander Held: [Physics analysis for the HL-LHC: concepts and pipelines in practice with the Analysis Grand Challenge](#), Tuesday 17:00
- Vincenzo Padulano: [First implementation and results of the Analysis Grand Challenge with a fully Pythonic RDataFrame](#), Tuesday 17:15
- ... and a lot more related to the topics we talked about this week! (coffea + dask-awkward, awkward, ServiceX, ROOT RDF, ...)

Thank you!

- To the **speakers** for preparing all the material
 - and those in the background making the talks & demos possible
- To the **local organizers**: Brian Bockelman, Kyle Cranmer, Matt Bialo
- To **all of you** for attending and contributing discussions

Stay in touch: analysis-grand-challenge@iris-hep.org (sign-up: [google group](#))

Have a safe trip!

And see some of you at CHEP!

