IRIS-HEP Project Proposal: Maxym Naumchyk

Improving and simplifying the Coffea library [1]

Problem setting:

- Coffea's schemas [2] are hard to maintain and develop
- Coffea does not support parquet data format currently
- Coffea can not easily support RNTuple data in the future
- Changes in dask-awkward and uproot can not easily propagate to Coffea.
- Historically, Coffea developed certain features ahead of dask-awkward and uproot. The features are now being introduced in these libraries and it is a good idea to make use of them directly.

Project layout:

- 3-layer approach:
  - I/O-layer: Make Coffea start with the output of `uproot.dask`, `uproot.open`, and `dask_awkward.from_parquet`. Output structure of these functions can be assumed to be the same.
  - Rearrange-layer: Rewrite Coffea's schema building by using "no-data-loading/touching" zipping operations [3] (similar to `ak.zip`) of the inputs coming from the I/O layer, instead of building awkward forms.
  - Schema-specific-layer: Propagate all schema specifics to the rearranged array (output of the "rearrange-layer"), e.g., particle collection cross-references for NanoAOD. This will allow to entirely remove some of the features implemented through a DSL [4]
- Implement the same 3 "layers" for all other schemas (Physlite, Delphes, …)
- Optional: Feedback and iteration: talk to users who do real analysis, help them migrate to the new version, implement missing features.

Project timeline and duration:

IO-layer + Rearrange-layer: ~4 weeks
Schema-specific-layer: ~2 weeks

Implementing the same layers for other schemas will be faster after finalizing the first schema, but requires communication with experts from the corresponding physics experiment: ~11 weeks

Improving the Physlite schema after implementing the 3-layer rearrangement: ~3 weeks

The project duration will be from February to July, for a total of 20 weeks.

Used links:

[1] https://github.com/scikit-hep/coffea
[2] https://github.com/scikit-hep/coffea/tree/master/src/coffea/nanoevents/schemas
[3] https://github.com/scikit-hep/awkward/blob/main/src/awkward/operations/ak_zip_no_broadcast.py
[4] https://github.com/scikit-hep/coffea/blob/master/src/coffea/nanoevents/transforms.py