Contribution ID: **31** Type: **Poster**

Describe Data to get Science-Data-Ready Tooling: Awkward as a Target for Kaitai Struct YAML

Monday, 11 March 2024 16:15 (30 minutes)

In some fields, scientific data formats differ across experiments due to specialized hardware and data acquisition systems. Researchers need to develop, document, and maintain specific analysis software to interact with these data formats. These software are often tightly coupled with a particular data format. This proliferation of custom data formats has been a prominent challenge for small to mid-scale experiments. The widespread adoption of ROOT has largely mitigated this problem for the Large Hadron Collider (LHC) experiments. However, not all experiments use ROOT for their data formats. Experiments such as Cryogenic Dark Matter Search (CDMS) continue to use custom data formats to meet specific research needs. Therefore, simplifying the process of converting a unique data format to analysis code still holds immense value for scientific communities even beyond HEP. We have added Awkward Arrays, a Scikit-HEP library for storing nested and variable data into Numpy-like arrays, as a target language for Kaitai Struct for this purpose.

Kaitai Struct is a declarative language that uses a YAML-like description of a binary data structure to generate the code to read a raw data file in any of the supported languages. Researchers can simply describe their custom data format in the Kaitai Struct YAML (KSY) language only once. The Kaitai Struct Compiler generates C++ code to fill the LayoutBuilder buffers using the KSY format. In a few simple steps, the Kaitai Struct Awkward Runtime API can convert the generated C++ code into a compiled Python module using ctypes. Finally, the raw data file can be passed to the module to produce Awkward Arrays.

This talk will introduce the Awkward Target for the Kaitai Struct Compiler and the Kaitai Struct Awkward Runtime API. It will demonstrate the conversion of a given KSY for a specific custom file format to Awkward Arrays.

Significance

Collaborations that use a custom data format spend many hours writing their own tools to read and analyse their data. It is difficult to fund such tools because they are not usable outside the collaboration, and as a result, the analysis can be restricted to what the original author envisioned. Even if the tool continues to have a maintainer, it is usually poorly documented and tested, making it difficult to continue maintaining it. It also poses a significant barrier for scientists entering the collaboration.

Switching to a supported standard data format sounds like the most obvious solution. However, in the case of the Cryogenic Dark Matter Search Collaboration, for example, this would require substantial rewriting of the data-acquisition system or switching to a different one. This would require additional personnel and substantial time investment. In addition, there are legacy datasets that still have the potential for new science. This project provides a simple and effective solution to this problem of reading and analysing custom data formats for small and mid-scale collaborations across the sciences. Instead of developing their own tools, the collaborations only need to describe their custom data formats in KSY language just once and then directly use the Kaitai Struct Awkward Runtime API to convert their data into Awkward Arrays.

References

Experiment context, if any

Cryogenic Dark Matter Search (CDMS)

Primary authors: GOYAL, Manasvi (Princeton University (US)); OSBORNE, Ianna (Princeton University); PIVARSKI, Jim (Princeton University); Dr ROBERTS, Amy (University of Colorado Denver); ZONCA, Andrea

Presenter: GOYAL, Manasvi (Princeton University (US))

Session Classification: Poster session with coffee break

Track Classification: Track 1: Computing Technology for Physics Research