PyHEP 2023 - ”Python in HEP” Users Workshop (online)

Contribution ID: **7** Type: **Tutorial**

Unified and semantic processing for heterogeneous Monte-Carlo data

Thursday 12 October 2023 14:00 (30 minutes)

This talk presents a comprehensive toolset tailored to high energy physics (HEP) research, comprising graphicle, showerpipe, colliderscope, and heparchy. Each component integrates into a cohesive workflow, addressing distinct stages of HEP data analysis.

These tools fill a niche within the HEP software ecosystem in Python by offering unified representations of Monte-Carlo event records. Routines are provided for filtering over all components of the event record, using operations based on the semantics of each particle attributes, *eg*. momentum, PDG code, status code, *etc*. These may be combined in complex queries, leading to powerful high level analysis.

showerpipe, as the foundational component, provides essential capabilities for simulating particle showers in HEP experiments. It provides a Pythonic interface with Pythia8 event generators and Les Houches files, generating simulated data. PythiaGenerator wraps pythia8 with a standard Python iterator. Successive events are generated by simply looping over an instance of this class. The event record is contained in a PythiaEvent, whose properties expose particle data via numpy arrays. This enables easy downstream analysis and portability in many pipelines. The ancestry of particles within the event is given as a directed acyclic graph, in coordinate (COO) format.

Building upon showerpipe‘s data generation capabilities, graphicle provides routines to unify the heterogeneous data record. It offers a sophisticated approach to semantic data querying. With graphicle, researchers can navigate ancestry via the DAG, select specific event regions based on status codes, form clusters, and extract aggregate properties, like mass. This semantic querying of heterogeneous data bridges the gap between data generation and analysis, streamlining the process.

colliderscope complements the toolchain by providing a dynamic visualization layer. It enhances data understanding with interactive HTML displays, including directed acyclic graphs (DAGs) for event history representation and scatter plots for particle distribution analysis. In addition to visualising individual event records, colliderscope offers histogram tools for statistical analysis of many events. colliderscope may be installed with an optional web-interface, which provides realtime data generation, visualisation, clustering, and mass calculations.

To complete the toolset, heparchy handles input and output (IO) operations for the generated data. It enables efficient data storage, retrieval, and management with HDF5 files. It standardises the difficult task of efficient storage of, and access to, HEP data. Data may be compressed either with LZF or GZIP, for wider compatibility outside of Python. heparchy may be integrated with dataloaders in ML toolchains to offer convenient and high performance data retrieval for training.

This talk will demonstrate how these libraries are used together to perform data analysis. Examples will show its use cases as an exploratory tool, a work-horse to produce high performance analysis scripts, or as part of a pre-processing pipeline for machine learning applications. Attendees will gain insights into the symbiotic relationship between data generation, semantic querying, visualization, and data management, resulting in a comprehensive toolset that empowers numerical scientists in HEP research.

Author: CHAPLAIS, Jacan

Presenter: CHAPLAIS, Jacan

Session Classification: Plenary Session Thursday