

Experiments Toward a Modern Analysis Environment: Using TMVA and other tools in a functional world with continuous integration for analysis

Thursday 13 October 2016 16:30 (15 minutes)

A modern high energy physics analysis code is complex. As it has for decades, it must handle high speed data I/O, corrections to physics objects applied at the last minute, and multi-pass scans to calculate corrections. An analysis has to accommodate multi-100 GB dataset sizes, multi-variate signal/background separation techniques, larger collaborative teams, and reproducibility and data preservation requirements. The result is often a series of scripts and separate programs stitched together by hand or automated by small driver programs scattered around an analysis team's working directory and disks. Worse, the code is often much harder to read and understand because most of it is dealing with these requirements, not with the physics. This paper describes a framework that is built around the functional and declarative features of the C# language and its Language Integrated Query (LINQ) extensions to declare an analysis. The framework uses language tools to convert the analysis into C++ and runs ROOT or PROOF as a backend to determine the results. This gives the analyzer the full power of an object-oriented programming language to put together the analysis and at the same time the speed of C++ for the analysis loop. A fluent interface has been created for TMVA to fit into this framework, and can be used as a model for incorporating other complex long-running processes into similar frameworks. A by-product of the design is the ability to cache results between runs, dramatically reducing the cost of adding one-more-plot. This lends the analysis to running on a continuous integration server after every check-in (Jenkins). To aid to data preservation a backend that accesses GRID datasets by name and transforms has been added as well. This paper will describe this framework in general terms along with the significant improvements described above.

Secondary Keyword (Optional)

Analysis tools and techniques

Primary Keyword (Mandatory)

Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional)

Author: WATTS, Gordon (University of Washington (US))

Presenter: WATTS, Gordon (University of Washington (US))

Session Classification: Posters B / Break

Track Classification: Track 5: Software Development