Contribution ID: **80** Type: **Oral**

Building a scalable python distribution for HEP data analysis

Tuesday 22 August 2017 17:25 (20 minutes)

There are numerous approaches to building analysis applications across the high-energy physics community. Among them are Python-based, or at least Python-driven, analysis workflows. We aim to ease the adoption of a Python-based analysis toolkit by making it easier for non-expert users to gain access to Python tools for scientific analysis. Experimental software distributions and individual user analysis have quite different requirements. One tends to worry most about stability, usability and reproducibility, while the other usually strives to be fast and nimble. We discuss how we built and now maintain a python distribution built for analysis while satisfying requirements both a large software distribution (in our case, that of CMSSW) and user, or laptop, level analysis. We pursued the integration of tools use in use by the broader data science community as well as HEP developed (e.g., histogrammar, root_numpy) Python packages. We discuss concepts we investigated for package integration and testing, as well as issues we encountered through this process. Distribution and platform support are important topics. We discuss our approach and progress towards a sustainable infrastructure for supporting this Python stack for the CMS user community and for the broader HEP user community.

Author: LANGE, David (Princeton University (US))

Presenter: LANGE, David (Princeton University (US))

Session Classification: Track 2: Data Analysis - Algorithms and Tools

Track Classification: Track 2: Data Analysis - Algorithms and Tools