Contribution ID: **215** Type: **Oral**

Big Data Analytics Tools as Applied to ATLAS Event Data

Thursday, 13 October 2016 15:00 (15 minutes)

Big Data technologies have proven to be very useful for storage, processing and visualization of derived metrics associated with ATLAS distributed computing (ADC) services. Log file data and database records, and metadata from a diversity of systems have been aggregated and indexed to create an analytics platform for ATLAS ADC operations analysis. Dashboards, wide area data access cost metrics, user analysis patterns, and resource utilization efficiency charts are produced flexibly through queries against a powerful analytics cluster. Here we explore whether these techniques and analytics ecosystem can be applied to add new modes of open, quick, and pervasive access to ATLAS event data so as to simplify access and broaden the reach of ATLAS public data to new communities of users. An ability to efficiently store, filter, search and deliver ATLAS data at the event and/or sub-event level in a widely supported format would enable or significantly simplify usage of big data, statistical and machine learning tools like Spark, Jupyter, R, SciPy,Caffe, TensorFlow, etc.. Machine learning challenges such as the Higgs Boson Machine Learning Challenge, the Tracking challenge, Event viewers (VP1, ATLANTIS, ATLASrift), and still to be developed educational and outreach tools would be able to access the data through a simple REST API. In this preliminary investigation we focus on derived xAOD data sets. These are much smaller than the primary xAODs having containers, variables, and events of interest to a particular analysis. Being encouraged with the performance of Elasticsearch for the ADC analytics platform, we developed an algorithm for indexing derived xAOD event data. We have made an appropriate document mapping and have imported a full set of standard model W/Z datasets. We compare the disk space efficiency of this approach to that of standard ROOT files, the performance in simple cut flow type of data analysis, and will present preliminary results on its scaling characteristics with different numbers of clients, query complexity, and size of the data retrieved.

Tertiary Keyword (Optional)

Analysi tools and techniques

Secondary Keyword (Optional)

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory)

Data model

Primary author: VUKOTIC, Ilija (University of Chicago (US))

Co-author: GARDNER JR, Robert William (University of Chicago (US))

Presenter: VUKOTIC, Ilija (University of Chicago (US))

Session Classification: Track 5: Software Development

Track Classification: Track 5: Software Development