Leveraging Public Clouds for DOE Environmental Streaming Data

Marty Humphrey
Associate Professor
Department of Computer Science
University of Virginia

As part of the team supporting the AmeriFlux (http://ameriflux.lbl.gov) and FLUXNET (http://fluxnet.fluxdata.org/) collaborations, we often have discussions regarding how best to reduce the barrier/processing/time that exists between the raw data in the field and the end scientists. While acknowledging the need to QA/QC and/or gap-fill, there is an overall desire to reduce/eliminate a reliance on sporadic/periodic FTP/SCP from Tower PIs. That is, we believe there is an opportunity to connect the data to the repository at LBL/ORNL via a streaming data model and mechanisms. However, there are many issues and software components that must be identified and created in other to facilitate this more direct connection, thereby reducing the "time to insight".

We have recently begun investigating and prototyping a software architecture that leverages Amazon Web Services (http://aws.amazon.com/) -- in particular Kinesis (https://aws.amazon.com/kinesis/) and Lambda (https://aws.amazon.com/lambda/) – as the basis for collecting and processing environmental data such as that for AmeriFlux and FLUXNET. Significant challenges include the management of a large number of sensors, identification/creation of the domain-specific functionalities necessary to layer upon base AWS capabilities, and large-scale cost estimation and management. Our goal/interest in participating in STREAM2016 is to discuss with mutually-interested parties the design of our system, and identify potential synergies with the broader research efforts of other STREAM2016 participants.