



Contribution ID: 1

Type: **not specified**

An Open Source Big Data Ecosystem

Tuesday 7 October 2014 09:30 (1 hour)

Big Data Challenges in several scientific domains including Astronomy with the Square Kilometre Array (SKA), Climate Science with the Intergovernmental Panel on Climate Change (IPCC) and intelligence projects with DARPA including XDATA and Memex. These challenges are of the Volume, Velocity and Variety mix (700TB/sec) from the SKA; 100s of thousands of files and formats in IPCC model to remote sensing data comparisons; language translation, automatic file identification of 50+ thousand files in the DARPA and other contexts).

To address these challenges I have proposed and published a Vision for Data Science in Nature that addresses these challenges through a combination of: (1) rapid science algorithm integration; (2) intelligent data movement; (3) automated and accurate extraction of text, metadata, and language from 1000s of file formats; and (4) the promotion of open source software and communities to push this agenda forward.

NASA, DARPA, NSF, and many government agencies in the US are seeing the benefits and reaping the reward of open source software products. From the traditional consumption model, to groups learning how to produce open source, and participate in community oriented ecosystems, there is a large emerging and fast paced environment that connects government to industry to academia and to the outside world.

In this talk, I will describe an Open Source Big Data ecosystem at the Apache Software Foundation and elsewhere including projects and progress towards implementing the Vision for Data Science.

Presenter: MATTMANN, Chris (NASA/JPL, USC)