Geophysical Research Abstracts
Vol. 20, EGU2018-11196, 2018
EGU General Assembly 2018

Cloud Based Hyperspectral Image Analysis via a Functional Data Model

Anne Wilson, Doug Lindholm, Tom Baltzer, and Chris Pankratz
Laboratory for Atmospheric and Space Physics, Boulder, United States (anne.wilson@lasp.colorado.edu)

Researchers in the Earth atmosphere remote sensing community need tools to effectively handle the hyperspectral data now becoming available. The massive data volumes of these datasets can demand an estimated 100 TB of storage for an analysis, a barrier to their use.

The Hylatis project addresses this problem by creating a toolset for hyperspectral image analysis in the cloud. Leveraging the Spark big data engine and Jupyter notebooks, the goal is to identify and develop operations for hyperspectral analysis, both general and project specific, while providing sufficient performance to do interactive analysis of very large volumes in a sharable, repeatable manner.

Via this toolset development, Hylatis explores the deeper informatics question of modeling scientific data. Because data are foundational in any analysis system, the way they are modeled, how they are represented in the software, impacts how easy or hard it is to perform arbitrary operations on them.

In particular, infusing scientific domain semantics at the data model level limits the datasets that can be integrated into a system due to a data model mismatch situation. This occurs particularly when integrating data from disparate domains. For example, many Earth data access and analysis tools provide the semantics of geolocated data, but do not have semantics and thus capabilities for integrating spectral data or data from another planet. These data model mismatches require increase coding effort and complexity to resolve, which in turn increases the project cost and potential for errors in translation.

Hylatis leverages the LaTiS data access middleware. LaTiS represents a dataset as a domain agnostic, algebraic function of independent and dependent variables. LaTiS uses this ‘functional data model’ as the core data representation. Scientific domain semantics are removed and reapplied at other layers in the LaTiS framework in a modular fashion. At the domain neutral ‘pure’ data layer, mathematic semantics are available for any computational purpose, supporting server side computations of any kind. As any dataset can be modeled this way, the problem of data model mismatch is avoided.

LaTiS is written in Scala and uses a functional programming style of coding. Of particular value to science, pure functional programming better supports both parallelization and proofs of correctness. Both of these qualities are growing in importance as data volumes grow and science is being challenged for its veracity.

Hylatis leverages the Spark big data engine to create operations on hyperspectral datasets. This amounts to creating a mapping from the functional data model to the relational data model used by Spark. The result will be a second interface to Spark based on the functional data model that will support the integration and analysis of various hyperspectral datasets.

This presentation will discuss Hylatis, the functional data model, and the value that functional programming can bring to science.