Geophysical Research Abstracts
Vol. 21, EGU2019-12075, 2019
EGU General Assembly 2019

The development of an Earth science cloud analytics reference architecture – lessons learned from Giovanni.

David Meyer (1), Chris Lynnes (1), Mahabaleshwa Hegde (2,1), Kwo-sen Kuo (3,1), Christine Smit (4,1), Angela Li (1), and Hailiang Zhang (2)

(1) NASA, Goddard Space Flight Center, United States (david.j.meyer@nasa.gov), (2) ADNET Systems, Inc., (3) ESSIC, University of Maryland, College Park, (4) Telophase Corporation

An ongoing challenge to Earth science research is the need to discover, understand, access and harmonize data from disparate and rapidly growing data sources for integrated analysis and visualization. These challenges can be addressed through the development of a service-based cloud architecture whose components include preprocessing, transformation, analysis and visualization services operating on analytics-optimized data stores (AODS). The AODS contain data that are optimized for various applications, and addressed in terms of commonly-accepted scientific variables scaled to standardized units. The architecture itself should support a framework upon which higher-level services and customized AODS can be readily developed.

This communication describes the history, current capabilities and future direction of Giovanni, the Geospatial Interactive Online Visualization and Analysis Interface environment, as an important case study addressing many of the challenges described above. Giovanni was originally developed for integrated analysis of Earth science data from a variety of observation platforms and Earth system models. The currently operational Giovanni allows for 22 different analytical and visualization services against over 2000 scientific variables archived at 6 different data centers, fulfilling thousands of user requests per day, and operating from a data "cache" containing data and metadata which have been harmonized for integrated analysis. Federated versions of Giovanni were deployed to distribute these capabilities across multiple data centers, from which important lessons were learned to develop a more robust analytics framework.

Giovanni is currently being refactored as "Giovanni-in-the-cloud" to address various components of a cloud analytics reference architecture. To date, highly modular prototypes have been built from common services within the Amazon Web Services environment, using data stores such as Apache Parquet, demonstrating dramatic improvements in performance over the operational version. Currently under development is an analytics pipeline whose standards-based components will include preprocessing, analysis and visualization services on cloud-optimized data stores on data from multiple sources, along with a notional provenance model. Discussed here are results to date, current developments and anticipated capabilities from this work.