Tackling the Four V’s with NEXUS

Frank Greguska, Kevin Gill, Thomas Huang, Joseph Jacob, Nga Quach, and Brian Wilson

Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109-8099, United States of America.

NASA’s Earth Observing System Data and Information System (EOSDIS) reports that over 15 petabytes (PB) of Earth observing information are archived among the 12 NASA Distributed Active Archive Centers (DAACs); with more being archived daily. The upcoming Surface Water & Ocean Topography (SWOT) mission is expected to generate about 26 PB of data in 3 years. NEXUS is a state of the art deep data analytic program developed at the Jet Propulsion Laboratory with the goal of providing near real-time analytic capabilities for this vast trove of data. Rather than develop analytic services on traditional file archives, NEXUS organizes data into tiles in order to provide a platform for horizontal computing. To provide near real-time analytic solutions for missions such as SWOT, a highly scalable data ingestion solution is developed to quickly bring data into NEXUS. In order to accomplish this formidable challenge, the “Four V’s” (Volume, Velocity, Veracity, and Variety) of Big Data must be considered.

NEXUS consists of an ingestion subsystem that handles the Volume of data by utilizing a generic tiling strategy that subsets a given dataset into smaller tiles. These tiles are then indexed by a search engine and stored in a NoSQL database for fast retrieval. In addition to handling the Volume of data being indexed, the NEXUS ingestion subsystem is built for horizontal scalability in order to manage the Velocity of incoming data. As the load on the system increases, the components of the ingestion subsystem can be scaled to provide more capacity. During ingestion, NEXUS also takes a unique approach to the Veracity and Variety of Earth observing information being ingested. By allowing the processing and tiling mechanisms to be customized for each dataset, the NEXUS ingest system can discard erroneous or missing data as well as adapt to the many different data structures and file formats that can be found in satellite observation data. This talk will focus on the functionality and architecture of the data ingestion subsystem that is a part of the NEXUS software architecture and how it relates to the Four V’s of Big Data.