Bringing Algorithms to Landsat Data Using Hadoop

Srinivas Yarlanki, Randolph Wynne, Valerie Thomas, A. Lynn Abbott

We have reached a point of data deluge with NASA's Earth Observing data streams and other 'big data' sources needed to better understand the complex dynamics of terrestrial ecosystems. The bottleneck for Earth Science research that requires multitemporal image analysis is data transfer and storage. At this time, users have to first download needed data from a repository and then process it using local institutional resources. By performing user-defined analysis at the repository itself, on the same computational infrastructure where the data is stored, the time and cost of performing such analysis can be improved by orders of magnitude. The users can then download end results of the analysis instead of raw data from the repository. Such a capability is provided by Hadoop, an open source distributed file system, which moves computation to data by sending the analysis code to the node where the data is stored. Since the computation is performed on the storage node itself, a Hadoop cluster has more throughput and smaller network traffic when compared to computing on a traditional HPC system or SAN / NAS systems. Hence unlike HPC / SAN / NAS systems it does not require a high-speed network to transfer data.

We are using a Hadoop repository for storing and enabling user defined processing of Landsat data and quantify savings resulting from the following: (1) Enabling comprehensive analyses and data-mining at the data repository without the need for large computational infrastructure at users' institutions. (2) Avoiding costly transmission of large data stores and allowing dynamic optimization of performance and resource utilization for a given computational task. (3) Reducing cost, size and risk associated with hosting and managing the data. (4) Enabling remote assembly of high performance computational or analytical workflows in an elastic resource environment.