Geophysical Research Abstracts
Vol. 19, EGU2017-8413-1, 2017
EGU General Assembly 2017

Quest for Value in Big Earth Data

Kwo-Sen Kuo (1,2,3), Amidu O Oloso (4,5), Mike L Rilee (6,7), Khoa Doan (8), Thomas L Clune (9), and Hongfeng Yu (10)

(1) NASA Goddard Space Flight Center, Greenbelt, Maryland, United States (kwo-sen.kuo@nasa.gov), (2) ESSIC, University of Maryland-College Park, Maryland, United States (kkuo@umd.edu), (3) Bayesics LLC, Bowie, Maryland, United States (kuo@bayesics.com), (4) NASA Goddard Space Flight Center, Greenbelt, Maryland, United States (amidu.o.oloso@nasa.gov), (5) SSAI Inc, Lanham, Maryland, United States, (6) NASA Goddard Space Flight Center, Greenbelt, Maryland, United States (michael.l.rilee-1@nasa.gov), (7) Rilee Systems Technologies, Derwood, Maryland, United States, (8) ESSIC, University of Maryland-College Park, Maryland, United States, (9) NASA Goddard Spacek Flight Center, Greenbelt, Maryland, United States (thomas.l.clune@nasa.gov), (10) University of Nebraska-Lincoln, Nebraska, United States (hfyu@unl.edu)

Among all the V's of Big Data challenges, such as Volume, Variety, Velocity, Veracity, etc., we believe Value is the ultimate determinant, because a system delivering better value has a competitive edge over others. Although it is not straightforward to assess the value of scientific endeavors, we believe the ratio of scientific productivity increase to investment is a reasonable measure. Our research in Big Data approaches to data-intensive analysis for Earth Science has yielded some insights, as well as evidences, as to how optimal value might be attained. The first insight is that we should avoid, as much as possible, moving data through connections with relatively low bandwidth. That is, we recognize that moving data is expensive, albeit inevitable. They must at least be moved from the storage device into computer main memory and then to CPU registers for computation. When data must be moved it is better to move them via relatively high-bandwidth connections and avoid low-bandwidth ones. For this reason, a technology that can best exploit data locality will have an advantage over others. Data locality is easy to achieve and exploit with only one dataset. With multiple datasets, data colocation becomes important in addition to data locality. However, the organization of datasets can only be co-located for certain types of analyses. It is impossible for them to be co-located for all analyses. Therefore, our second insight is that we need to co-locate the datasets for the most commonly used analyses. In Earth Science, we believe the most common analysis requirement is "spatiotemporal coincidence". For example, when we analyze precipitation systems, we often would like to know the environment conditions "where and when" (i.e. at the same location and time) there is precipitation. This "where and when" indicates the "spatiotemporal coincidence" requirement. Thus, an associated insight is that datasets need to be partitioned per the physical dimensions, i.e. space and time, rather than their array index dimensions to achieve co-location for spatiotemporal coincidence. This leads further to the insight that, in terms of optimizing Value, achieving good scalability in Variety is more crucial than good scalability in Volume. Therefore, we will discuss our innovative approach to improving productivity by homogenizing the daunting varieties in Earth Science data to enable data co-location systematically. In addition, a Big Data system incorporating the capabilities described above has the potential to drastically shorten the data preparation period of machine learning, better facilitate automated machine learning operations, and further boost scientific productivity.