EGU2020-20339
https://doi.org/10.5194/egusphere-egu2020-20339
EGU General Assembly 2020

Analytics Optimized Geoscience Data Store with STARE-based Packaging

Kwo-Sen Kuo[1] and Michael Rilee[2]
[1]Bayesics LLC, Bowie, Maryland, USA (kuo@bayesics.com)
[2]Rilee Systems Technologies, LLC, Derwood, Maryland, USA (mike@rilee.net)

The only effective strategy to address the volume challenge of Big Data is “parallel processing”, e.g. employing a cluster of computers (nodes), in which a large volume of data is partitioned and distributed to the cluster nodes. Each of the cluster nodes processes a small portion of the whole volume. The nodes, working in tandem, can therefore collectively process the entire volume within a much-reduced period of time. In the presence of data variety, however, it is no longer as straightforward, because naïve partition and distribution of diverse geo-datasets (packaged with existing practice) inevitably results in misalignment of data for the analysis. Expensive cross-node communication, which is also a form of data movement, thus becomes necessary to bring the data in alignment first before analysis may commence.

Geoscience analysis predominantly requires spatiotemporal alignment of diverse data. For example, we often need to compare observations acquired by different means & platforms and compare model output with observations. Such comparisons are meaningful only if data values for the same space and time are compared. With the existing practice of packaging data using the conventional array data structure, it is nearly impossible to spatiotemporally align diverse data. Because, while array indices are generally used for partition and distribution, for different datasets (even data granules) the same indices most-often-than-not refer to different spatiotemporal neighborhoods. Partition and distribution using conventional array indices thus often results in data of the same spatiotemporal neighborhoods (from different datasets) reside on different nodes. Comparison thus cannot be performed until they are brought together to the same node.

Therefore, we need indices that tie directly and consistently to spatiotemporal neighborhoods to be used for partition and distribution. SpatioTemporal Adaptive-Resolution Encoding (STARE) provides exactly such indices, which can replace floating-point encoding of longitude-latitude and time as a more analytics-optimized alternative. Moreover, data packaging can base on STARE indices. Due to its hierarchical nature, geo-spatiotemporal data packaged based on STARE hierarchy offers essentially a reusable partition for distribution adaptable to various computing-and-storage architectures, through which spatiotemporal alignment of geo-data from diverse

sources can be readily and scalably achieved to optimize parallel analytic operations.