Abstract: For decades, geoscience data practice has followed a two-step approach: 1) package data into files for archival and distribution and 2) catalog metadata of the data files into databases managed by relational database management systems, making them discoverable and searchable. The establishment of popular distributed active archive centers and the standardization of file formats, such as HDF/netCDF, have become a de facto standard used by domestic/international organizations. Despite its success, this 2-step approach has reached its limit with today’s demands. It inevitably leads to data transfer, mostly via low-bandwidth Internet connections, for processing or integration by users. The problem is further compounded by the enormous variety of geoscience data that arises from varying geometries and resolutions used for model simulations and, even more, from the different modes, resolutions, and geometries of observations. These lead to poor volume and variety scaling.

Among the solutions, loosely coupled and tightly coupled categories of approaches (according to the strength of coupling between storage and memory, and thus compute) are regarded as better suited for I/O-bound scenarios. With the loosely coupled approaches, exemplified by e.g. Spark and Hadoop, the analysis engines are loosely coupled with storage systems and have no access to the detailed map of partition locations; the analysis engines in tightly coupled approaches, exemplified by parallel distributed database management systems, e.g. SciDB, however, do. In this poster, we present evaluations of representative approaches from these two categories. We show that the distribution of data traffics in loosely coupled systems skews towards network bandwidth, while that of the tightly coupled systems relies more on the effectively higher storage bandwidth. We develop new distributed data placement and analysis techniques and demonstrate that a tightly coupled approach holds the promise to best exploit the data movement bandwidths and more comprehensively meet geoscience Big Data challenges.

Figure 1. Comparison between traditional data archive centers, loosely coupled approaches, and tightly coupled approaches: Data archive centers and loosely coupled approaches involve high data movement traffic (orange arrows). Tightly coupled approaches provide a more scalable solution for reducing data movement (green arrows) and providing optimal data analysis efficiency and value.