Thesis: Scalable Spatial Partitioning.

Abstract: Distributed spatial query processing over large datasets has become an essential step in every spatial data management system. General-purpose cluster computing frameworks like Apache Spark do not differentiate between spatial and non-spatial data and makes them ill-equipped for efficient spatial query processing. Spatial extensions utilize the capabilities of cluster computing frameworks and add spatial data support through spatial object recognition and spatial partitioning. Partitioning improves query processing by spatially grouping objects on processing nodes, maximizes resource utilization. Several problems that affect the scalability of spatial partitioning include minimizing query skews, accounting for boundary-crossing objects, and minimizing I/O intensive operations. In this work, we propose a novel approach for producing a fast and scalable spatial partitioning scheme which guides the execution of distributed spatial queries. We detail the obstacles that face big data partitioning and propose a set of solutions that collectively produce a scalable and reusable spatial partitioning scheme. The scheme customizes itself based on the traits of the input datasets and accounts for non-spatial data, the objects' types and distribution, and the dataset's size. We implement our solution, apply it to the well-known kNN query, and compare its speed and accuracy to existing techniques. Preliminary experiments show that the kNN query with the spatial partitioner, on average, produce near accurate results (99% – 100%) in approximately half of the time.

Committee:

- Professor Huy T. Vo, The Cit College Of New York
- Professor Feng Gu, The College Of Staten Island
- Professor Kaliappa Ravindran, The City College Of New York

Outside member:

- Professor Kai Zhao, Robinson College Of Business, Georgia State University