Dissertation Defense

High Performance Spatial Query Processing for Large Scale Spatial Data Warehousing

Ablimit Aji
Emory University

Abstract: Support of high performance queries on large volumes of spatial data have become important in many application domains, including geowspatial problems in numerous fields, location based services, geo-social networks, and emerging scientific applications that are increasingly data- and compute-intensive. There are two major challenges for managing and querying massive spatial data: the explosion of spatial data, and the high computational complexity of spatial queries due to the multi-dimensional nature of spatial analytics. High performance computing capabilities are fundamental to efficiently handling of massive spatial datasets. MapReduce based computing model provides a highly scalable, reliable, elastic and cost effective framework for processing massive data on a cluster or cloud environment. While the MapReduce model fits nicely with large scale problems through data partitioning, spatial queries and analytics are intrinsically complex to fit into the MapReduce model easily. Meanwhile, hybrid systems combining CPUs and GPUs are becoming commonly available in commodity clusters, but the computing capacity of such systems is often underutilized. Providing new spatial querying and analytical methods to run on such architecture requires us to answer several fundamental research questions that are of practical importance. The goal of my dissertation is to create a framework with new systematic methods to support high performance spatial queries for spatial big data on MapReduce and CPU-GPU hybrid platforms, driven by real-world use cases. Towards that end, we have researched multi-level parallelism methods of spatial queries running on these platforms. Specifically, we have conducted following studies: 1) create new spatial data processing methods and pipelines with spatial partition level parallelism through a simple programming model MapReduce and propose multi-level indexing methods to accelerate spatial data processing, 2) develop two critical components to enable data parallelism: effective and scalable spatial partitioning in MapReduce (pre-processing), and query normalization methods for partition effect, 3) integrate GPU-based spatial operations into MapReduce pipelines 4) investigate optimization methods for data skew mitigation, and CPU/GPU resource coordination in MapReduce, and 5) support declarative spatial queries for workload composition, and create a query translator to automatically translate the queries into MapReduce programs. Consequently, we have developed Hadoop-GISb a MapReduce based high performance spatial querying system for spatial data warehousing. The system supports multiple types of spatial queries on MapReduce through spatial partitioning, implicit parallel spatial query execution on MapReduce, and effective methods for amending query results through handling bound- ary objects. Hadoop-GIS utilizes global partition indexing and customizable on demand local spatial indexing to achieve efficient query processing. Hadoop-GIS is integrated into Hive to support declarative spatial queries with an integrated architecture. The systems and developed approaches are released as an open source software package for use.

Friday, October 31, 2014, 3:00 pm
Mathematics and Science Center: W303

Advisor: Fusheng Wang

MATHEMATICS AND COMPUTER SCIENCE
EMORY UNIVERSITY