Stony Brook University

Hadoop-GIS
Spatial Big Data Solutions

Scalable 3D Spatial Queries for Analytical Pathology Imaging with MapReduce

Yanhui Liang[1], Hoang Vo[1], Ablimit Aji[2], Jun Kong[3], Fusheng Wang[1]
[1]Department of Computer Science, Stony Brook University, [2]Hewlett Packard Labs, [3]Department of Biomedical Informatics, Emory University

Introduction

Background

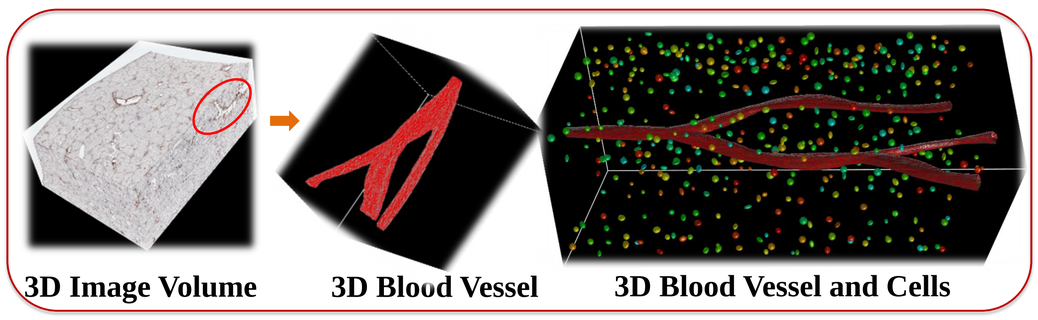
- Massive amount of 3D spatial objects derived from 3D analytical pathology imaging
- "GIS" generic queries for pathologically meaningful spatial analysis on 3D big data
  e.g., for each 3D cell, return the nearest 3D blood vessel and the distance

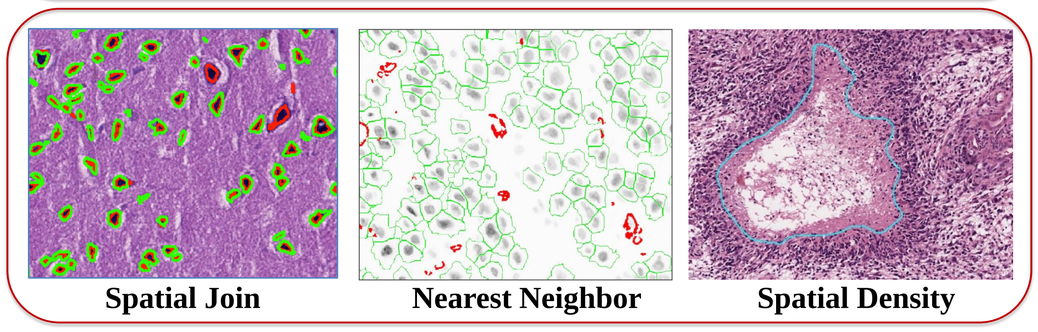
Challenges

- Explosion of 3D data: hundreds of millions of 3D objects in the scale of terabytes
- Complex structures and representations: 3D mesh model with numerous structural details
- High computation complexity: computationally intensive 3D geometric operations

Objectives

- To provide a scalable and efficient 3D spatial query system for querying massive 3D spatial data based on MapReduce





Scalable 3D Spatial Queries with MapReduce

System Overview

- Spatial data partitioning: partition 3D input into cuboids to increase the level of parallelism
- 3D on-demand spatial query engine: multi-level indexing and spatial query processors

3D Spatial Partitioning

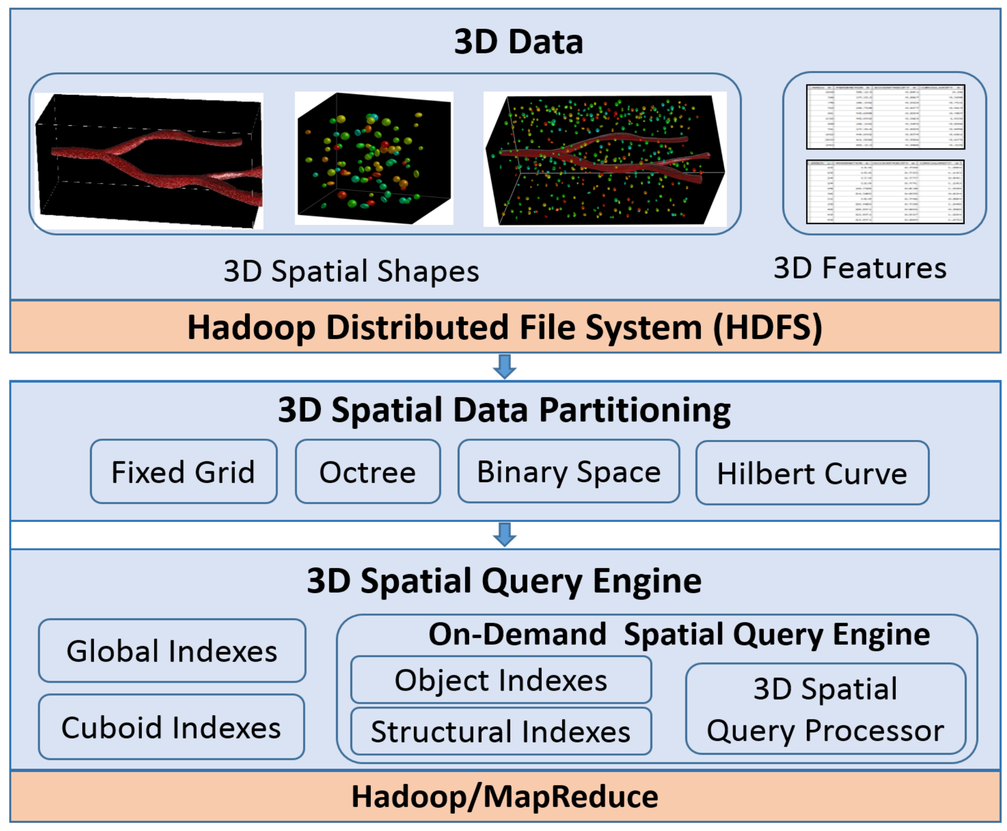
- Each partitioned cuboids as a processing unit for querying tasks on MapReduce
- Cuboids are processed in parallel without data dependence or communication requirements

Multi-level Indexing

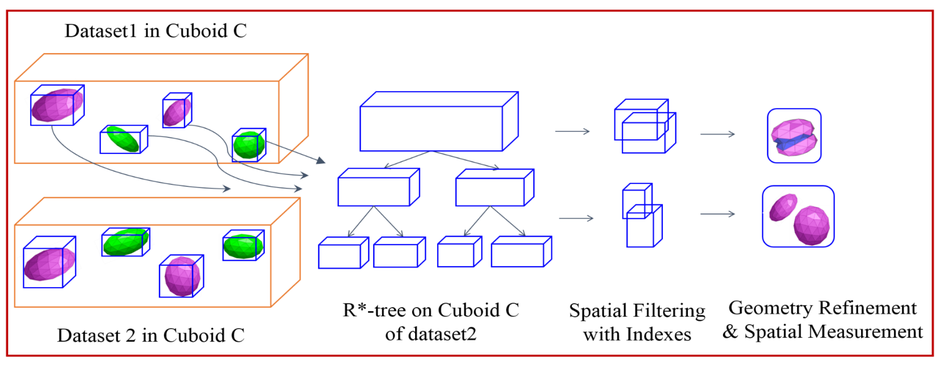
- Global storage indexing: HDFS-level data filtering (point query)
- Cuboid indexing: MapReduce-level computation filtering (containment query)
- On-demand object indexing: to build indexes on-the-fly for objects within a cuboid
- Structural indexing: for geometric computation on 3D objects with complex structures

3D Spatial Queries

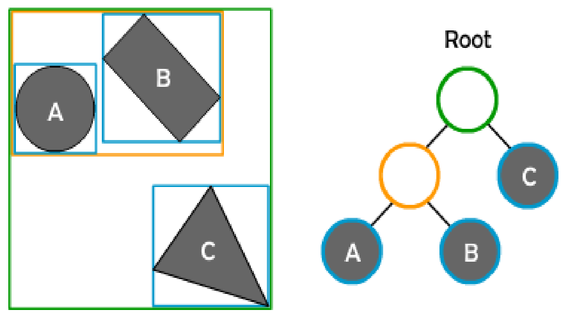
- Spatial join or spatial cross-matching
- Nearest neighbor: 3D R*-tree (with *AABB* tree) and Voronoi diagram (with skeleton)



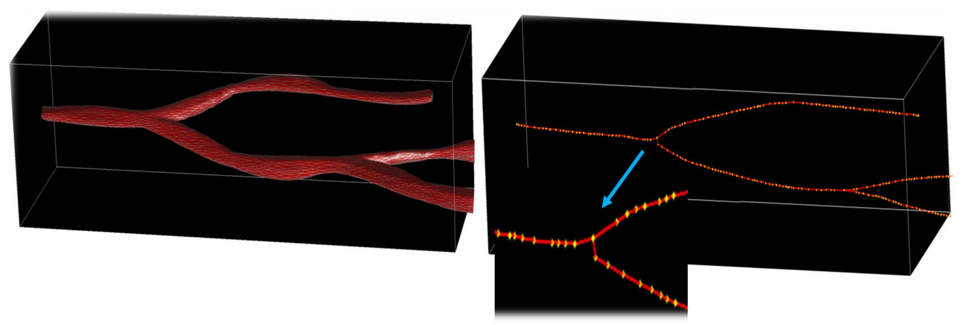
System Architecture of Hadoop-GIS 3D



The Workflow of 3D Spatial Join



AABB Tree Illustration with 2D Data



Structure Index of 3D Blood Vessel with Skeleton

Experimental Results

Performance of 3D Spatial Queries

- The cluster has five nodes with 124 cores in total; Each node has 5TB hard drive and 128GB memory
- Five datasets with millions of 3D cells, and several thousand of 3D blood vessels
- Three 3D datasets for Level of Details (LoD) testing

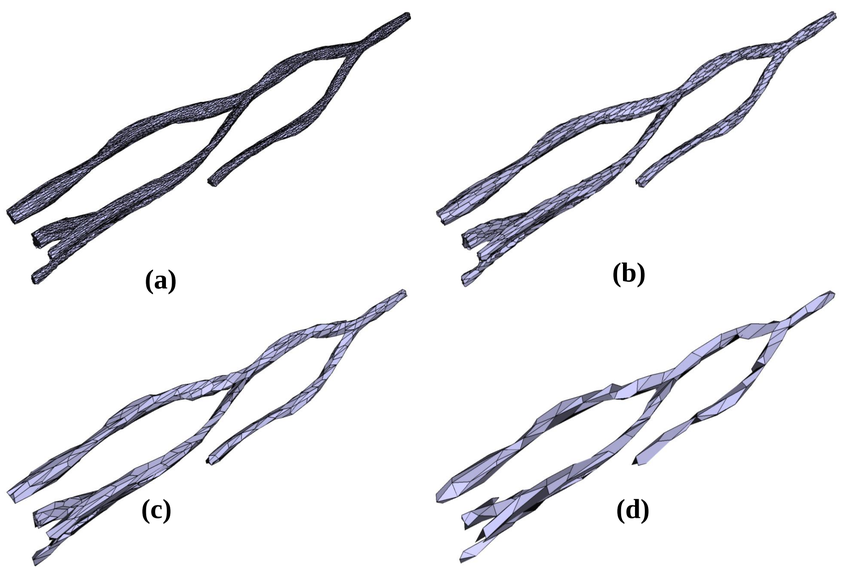
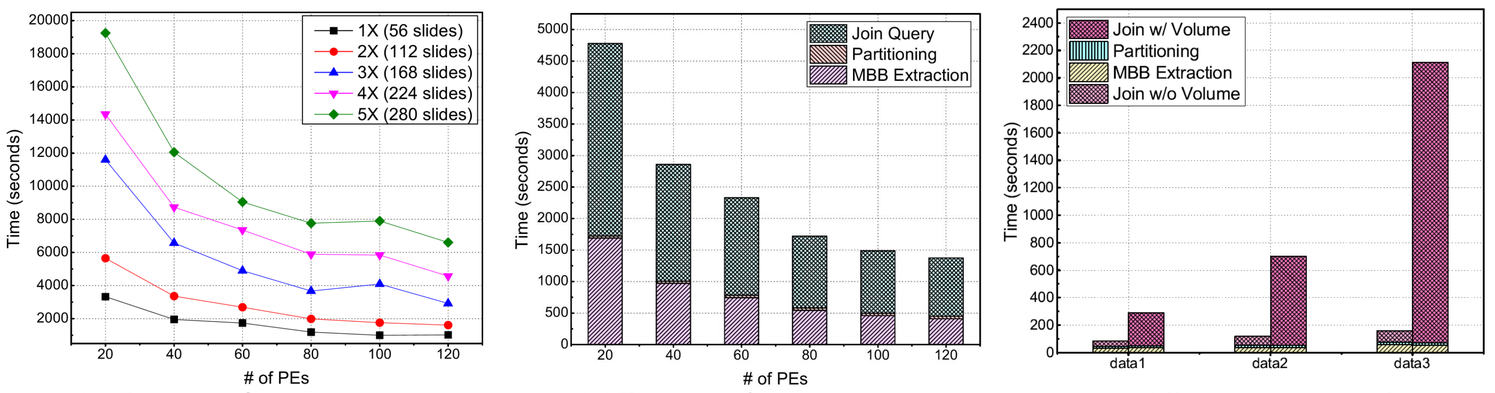
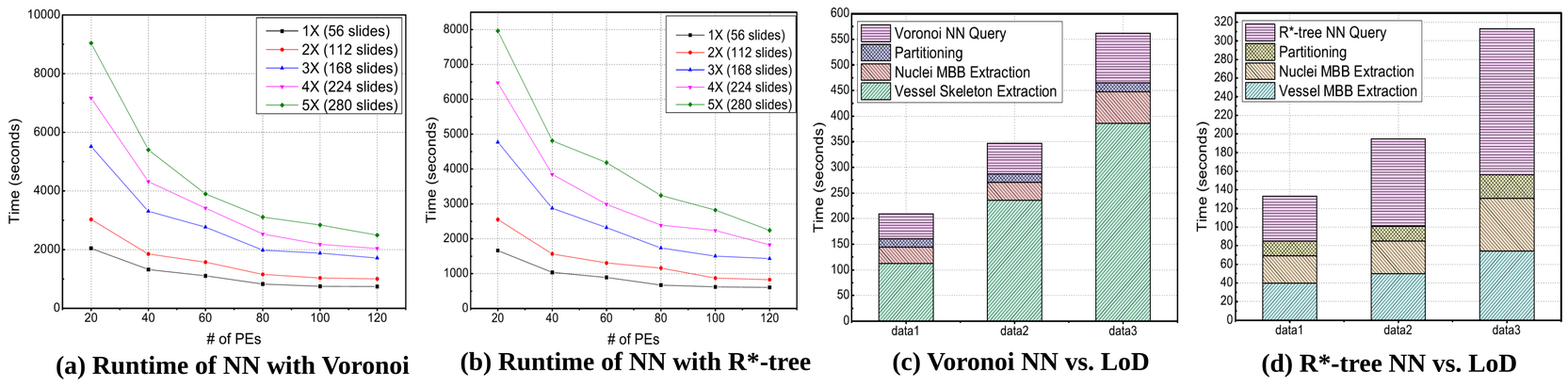


Fig. 2: Different LoD Representations for the Same Blood Vessel



(a) Runtime of spatial join (b) Runtime of each component (c) Spatial join vs. LoD

Fig. 1: System Performance of Spatial Join



(a) Runtime of NN with Voronoi (b) Runtime of NN with R*-tree (c) Voronoi NN vs. LoD (d) R*-tree NN vs. LoD

Fig. 3: System Performance of Nearest Neighbor (NN) Query

Acknowledgements

This research is supported in part by grants from National Science Foundation ACI 1443054 and IIS 1350885, National Institute of Health K25CA181503, the Emory University Research Committee, Pitney Bowes, Amazon and Google.