CIF21 DIBBs: Scalable Capabilities for Spatial Data Synthesis (NSF 1443080)

PI: Shaowen Wang[1,2,3] Co-PIs: Kate Keahey[4] and Anand Padmanabhan[1,2,3]

[1]CyberGIS Center for Advanced Digital and Spatial Studies
[2]Department of Geography and Geographic Information Science
[3]National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
[4]Argonne National Laboratory and University of Chicago
shaowen@illinois.edu, keahey@mcs.anl.gov, apadmana@illinois.edu

Objectives

1. Develop a core set of community-driven and scalable capabilities for meeting the requirements of spatial data synthesis in representative scientific case studies
2. Establish a suite of scalable data integration and aggregation capabilities by leveraging cyberGIS – geographic information science and systems (GIS) based on advanced cyberinfrastructure
3. Evaluate and improve these capabilities by engaging the broad cyberGIS community that spans bio, computational, engineering, geo, and social sciences
4. Integrate the data synthesis capabilities with the CyberGIS Science Gateway to ensure open and broad access to the capabilities
5. Develop novel education and training materials for a large number of users to learn the capabilities and related scientific principles of spatial data synthesis

Architecture

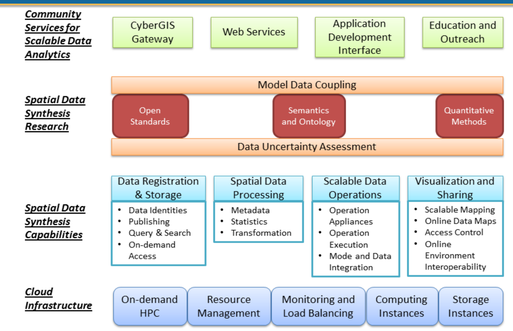


Figure 1. Architecture of scalable spatial data synthesis

Spatial Data Synthesis Capabilities

1. **Spatial data processing** capabilities for processing a wide range of spatial big data streams and types (e.g., spatial clustering for vector data and projection and resampling for raster data)
2. **Spatial data integration** capabilities for integrating location-based social media data with authoritative geospatial data sources
3. **Spatial data presentation** capabilities for interactively visualizing complex spatial relations (e.g. visualizing geospatial flows and spatial interactions) and uncertainties associated with geospatial big data
4. **Data retrieval and storage** capabilities for efficient and scalable storage of geospatial big data (e.g. fast geohash-based indexing (Figure 2) for handling rich geometries and varying spatial resolutions)

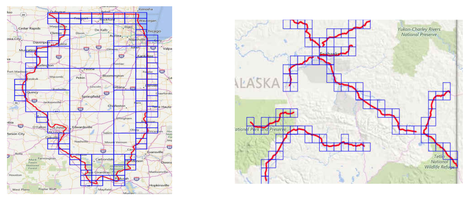
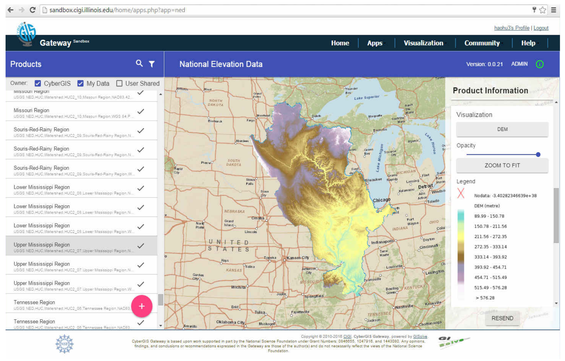


Figure 2. Indexing spatial objects using geohash: (a). Polygon indexing; (b). Line indexing

Case Study 1: TopoLens

- Prototyped a cyberGIS-based application for spatial big data synthesis and sharing by leveraging heterogeneous cyberinfrastructure with both cloud and HPC resources
- Employed modern microservices architecture
- Lowered barriers to accessing, processing, analyzing, and visualizing large raster datasets
- Data from 18 hydrological regions (Hydrological Unit Code level 2) and 48 conterminous states available



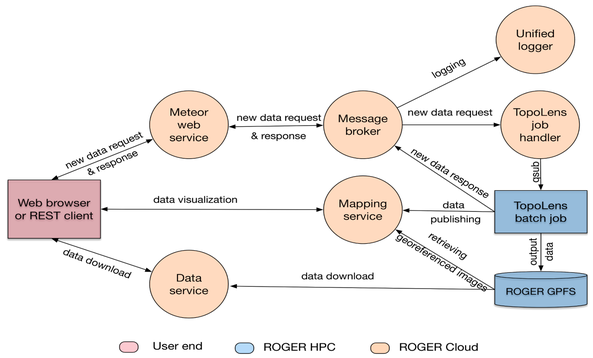
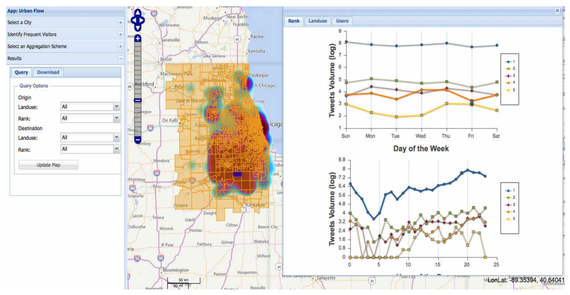


Figure 3. (a). TopoLens user interface, the selected DEM data is for the Upper Mississippi Region in NAD 83/Conus Albers projection; (b). TopoLens computation workflow expressed using microservices architecture
TopoLens is currently available at http://sandbox.cigi.illinois.edu.

Case Study 2: UrbanFlow

- Synthesized geospatial big data from social media with authoritative land use data
- Investigated population dynamics and land use changes based on the synthesized data
- Implemented a cyberGIS workflow using Apache Hadoop to process large streaming data from Twitter (e.g. 965 million geo-tagged tweets in 2014)
- Developed a novel distributed point-in-polygon algorithm suitable for large vector datasets



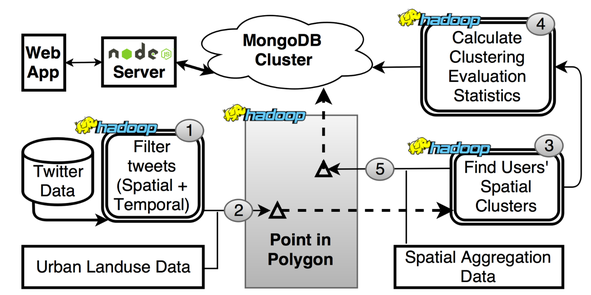


Figure 4. (a). UrbanFlow user interface; (b). CyberGIS workflow

Geospatial Operation Management

Our publishing and execution platform allows users to (1) publish versioned operations and appliances for computational reproducibility, (2) reference the exact operations used to obtain scientific results easily, (3) support execution on diverse platforms (e.g. XSEDE, NSF clouds, commercial platforms), and (4) control response time via elastic scaling to cloud resources.

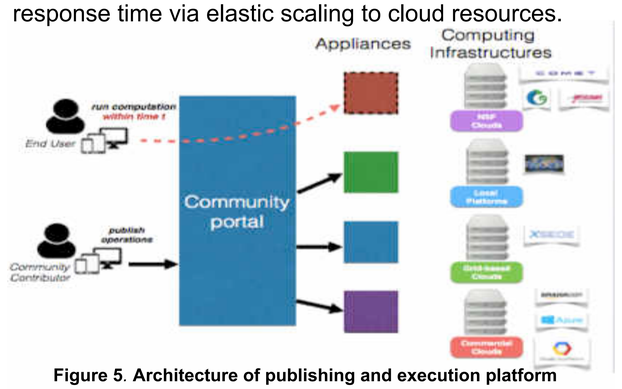


Figure 5. Architecture of publishing and execution platform

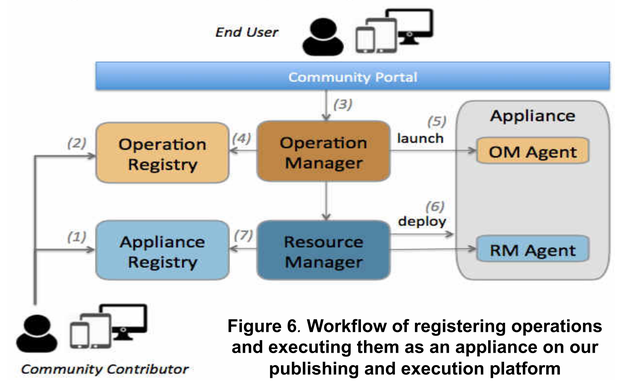


Figure 6. Workflow of registering operations and executing them as an appliance on our publishing and execution platform

Education and Training

- CyberGIS Summer School and a series of training workshops conducted at the University of Illinois at Urbana-Champaign were attended by over 300 participants who benefited from hands-on experience with the spatial data synthesis capabilities.
- The project team organized the first NSF Geospatial Data Science Workshop held in conjunction with CyberGIS 2016 on July 25-26 @ Urbana, Illinois, USA.

Contact: Shaowen Wang (shaowen@illinois.edu)

Acknowledgements: These materials are based upon work supported in part by the National Science Foundation under Grant No. 1443080

CyberGIS Center for Advanced Digital and Spatial Studies
http://www.cybergis.illinois.edu/