CyberGIS-Jupyter for Reproducible Geospatial Research and Education at Scale

Anand Padmanabhan

CyberGIS Center for Advanced Digital and Spatial Studies
Department of Geography and Geographic Information Science
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Location Powers: Data Science

November 13, 2019
Can we have a single tool that covers all the lifecycle of a scientific idea, from data to publication?
RESEARCH ARTICLE

CyberGIS-Jupyter for reproducible and scalable geospatial analytics

Dandong Yin¹,²  |  Yan Liu¹,²,³  |  Hao Hu¹,²,³  |  Jeff Terstriep³  |  Xingchen Hong³  |  Anand Padmanabhan¹,²  |  Shaowen Wang¹,²,³

¹Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Illinois
²CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Illinois
³National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Illinois

Correspondence
Shaowen Wang, Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Illinois; or CyberGIS Center for Advanced Digital and Spatial Studies, University of Illinois at Urbana-Champaign, Illinois.
Email: shaowen@illinois.edu

Present Address
Department of Geography and Geographic Information Science, Natural History Building, 1301 W Green Street, Champaign, IL 61801, USA

Summary
The interdisciplinary field of cyberGIS (geographic information science and systems (GIS) based on advanced cyberinfrastructure) has a major focus on data- and computation-intensive geospatial analytics. The rapidly growing needs across many application and science domains for such analytics based on disparate geospatial big data poses significant challenges to conventional GIS approaches. This paper describes CyberGIS-Jupyter, an innovative cyberGIS framework for achieving data-intensive, reproducible, and scalable geospatial analytics using Jupyter Notebook based on ROGER, the first cyberGIS supercomputer. The framework adapts the Notebook with built-in cyberGIS capabilities to accelerate gateway application development and sharing while associated data, analytics, and workflow runtime environments are encapsulated into application packages that can be elastically reproduced through cloud-computing approaches. As a desirable outcome, data-intensive and scalable geospatial analytics can be efficiently developed and improved and seamlessly reproduced among multidisciplinary users in a novel cyberGIS science gateway environment.

KEYWORDS
cloud computing, computational reproducibility, cyberGIS, geospatial big data

Bridging Local Cyberinfrastructure and XSEDE with CyberGIS-Jupyter

[Extended Abstract]

Anand Padmanabhan
Department of Geography and Geogic Information Science
University of Illinois of Urbana-Champaign
apadmaga@illinois.edu

Dandong Yin
Department of Geography and Geogic Information Science
University of Illinois of Urbana-Champaign
dyin4@illinois.edu

Fangzheng Lyu
Department of Geography and Geogic Information Science
University of Illinois of Urbana-Champaign
flu8@illinois.edu

Shaowen Wang
Department of Geography and Geogic Information Science
University of Illinois of Urbana-Champaign
shaowen@illinois.edu

INTRODUCTION

The fabric of national and international cyberinfrastructure ecosystems for scientific discovery and innovation can be viewed as distributed computing environments composed of powerful supercomputers, various cloud computing resources, and numerous local cyberinfrastructure (including both cloud and HPC) resources. However, extensive computational work conducted by academic researchers are often siloed in one of these environments. Science Gateways [5, 6], by simplifying access to advanced cyberinfrastructure resources, have made significant progress on connecting these silos by enabling researchers in many fields to access advanced cyberinfrastructure through web browsers. In this context, this research bridges between national and local cyberinfrastructure resources through: (a) horizontal scaling of CyberGIS-Jupyter between the cloud resources provided by JetStream on the Extreme Science and Engineering Discovery Environment (XSEDE) and a VMWare-based cloud environment on Virtual ROGER, a local cyberinfrastructure resource hosted by the CyberGIS Center for Advanced Digital and Spatial Studies at the University of Illinois at Urbana-Champaign campus; and (b) enabling the submission of computationally intensive models to the batch systems of both Virtual ROGER and Comet (an XSEDE resource). Specifically, we have developed a mechanism to provide access to a scalable JupyterHub platform together with cutting-edge cyberGIS software and hardware [4], called CyberGIS-Jupyter [7], which allows seamless access to high-performance computing (HPC) resources while shielding the complexity of managing cyberinfrastructure access from users. The user-friendly environment provided by CyberGIS-Jupyter along with computational scalability achieved through this research provides a powerful environment for conducting collaborative and reproducible research at scale with seamless access to advanced cyberinfrastructure at both local and national levels. The rest of this paper describes the architecture of our solution and articulates the corresponding implementation.

ARCHITECTURE AND IMPLEMENTATION

The architecture consists of four major layers: (a) user layer; (b) application layer; (c) cloud resources; and (d) HPC resources. The components of each of these layers and the interactions between them can be depicted in Figure 1. There are two levels at which users interact with cyberinfrastructure resources: (1) logging into JupyterHub and accessing their single user Jupyter Notebook server as an interactive session which runs transparently on local and national resources, and (2) submitting computationally intensive jobs that leverage HPC resources locally and on XSEDE.

CyberGIS

- CyberGIS -- geographic information science and systems (GIS) based on advanced computing and cyberinfrastructure
  - New-generation GIS
  - Focus on computational and data-intensive geospatial problem-solving within various research and education domains
  - Bridge gaps between geospatial big data, software and applications through advanced cyberinfrastructure

Features

• Achieves data-intensive, reproducible, and scalable geospatial analytics using Jupyter Notebooks
  – Provides a holistic solution
  – Makes sharing codes and workflows easy

• Reduces the barrier to accessing advanced cyberinfrastructure and cyberGIS capabilities
  – Exploits JupyterHub, cloud and high performance computing resources
Capabilities

• Provides notebook servers with cyberGIS libraries and many geospatial software packages installed
  — Built-in cyberGIS capabilities to accelerate gateway application development
    • E.g., HAND Application
  — Data, analytics, and workflow runtime environments are encapsulated into application packages

• Deployment can be elastically scaled to accommodate the computational needs of cyberGIS users
  — Straightforward management and maintenance of computational infrastructure
  — Seamless scaling between Virtual ROGER and XSEDE JetStream
Geovisualization

- Interactive map generation inside notebooks
- Support multiple formats of geospatial data
- Layer management, transparency and styles
Technologies

Batch Job management

Cloud-based architecture

Data/storage synchronization
Containers might run on any of the VMs. Shared file system allows data to be persisted between sessions.

HPC jobs to submitted from the Jupyter Notebook to batch system on virtual ROGER and Comet using CyberGIS libraries.
Case study

● Height Above Nearest Drainage (HAND) at 10m for continental US
  ○ Flood analysis map derived from 10m USGS 3DEP national elevation dataset (180 billion cells) and National Hydrography Dataset (2.67 million stream reaches, raw data size: 5.2TB)

● Goals
  ○ Collaborative methodology development
  ○ Scalable data analytics
  ○ Deliver methodology and data products to different user communities
    ■ Collaborators
    ■ Researchers
    ■ Decision makers
    ■ Students
Map Flood Inundation at Continental Scale

1. Data preparation and pre-processing

- Fetch the hands-on example data, from GIS in Water Resources (USU CEE5440, UT Austin CE 394K.3)

- The study area is Onion Creek in Texas. This watershed is a HUC 10 unit with code 12090120004.

- We use CyberGIS accelerated TauDEM version 28 and disflow direction calculation.

- More about TauDEM

- USGS 30E9P Elevation dataset (aka. National Elevation Dataset) is deployed on ROGIB as a VRT raster (whole dataset contains 477GB GeoTIFF, 1.86 billion onlin).

- NHDPlus2 (2.57 million flows in total) and associated water boundary dataset (WBD) are in Esri FileGDB format.

2. Workflow for computing HAND

- Find inlets

  - Use the dangle operation to find channel heads (inlets) of streams. They will guide the DEM-based stream calc.

- Rasterize inlet points to the same spatial extent as the input DEM. The output is the weight grid.

- To visualize a GeoTIFF, we need to:
  1. Render color from values;
  2. Making tiles (pyramids) from the colored image;

2.2 Fitremove

- TauDEM fitremove: hydro-confound the DEM

- Output is the TauDEM fitremove output.dat file for the data.

- Example version 1.8.3

- Some TauDEM functions All the necessary conditional tests.

- For this example, we find key points using optical images and orthoimages:

- Key points:

- log-log of ground truth data in line.dat.

3. Application

- Whose homes will be flooded should a 5m flood take place?

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases

- More about valid calcium and oil well databases
AAG-UCGIS Summer School 2019

- Week-long summer school on Reproducible Problem Solving with CyberGIS and Geospatial Data Science
- Held between July 8 and 13, 2019 at CyberGIS Center at UIUC
- Students created notebooks collaboratively to solve geospatial problems
- Over 35 students and 10 mentors engaged for intensive problem solving using CyberGIS-Jupyter
- Over 50 Jupyter notebooks were simultaneous running as docker container using a distributed computing infrastructure

Summer School students presenting their work using CyberGIS-Jupyter
GOAL

Conceptualize a Geospatial Software Institute (GSI) as a long-term hub of excellence in software infrastructure that can serve diverse research and education communities.
Acknowledgments

- National Science Foundation
  - ACI-1443080
  - ACI-1429699
  - OAC-1664119
  - OAC-1743184
  - XSEDE
  - XSEDE ECSS
Questions?

Contact

apadmana@Illinois.edu