Big Geospatial Analytics on the Cloud

Dinh Tran, Er Ran Khoo and Varun Chandola
Computer Science and Engineering
University at Buffalo
WorldWind for Scientific Community

- Frameworks to build geospatial applications
- Ability to do it in a browser is an additional bonus
- Excellent for many, including analysts, educators, policymakers
- What about the scientific community?
What does the Scientific Community Need?

1. Ability to **handle** scientific data
   - NetCDF, HDF

2. Scale to massive **data sizes**
   - For visualization and analytics
A Simple Use Case

- **Problem**: Study the impact of climate change on a particular location on the globe

Step 1: Get Data  
Step 2: Analyze  
Step 3: Visualize
webGlobe Capabilities

User selects a location
webGlobe Capabilities

Enters location of climate data
webGlobe Capabilities

Submits Request
webGlobe Capabilities
What happens under the hood?

- webGlobe dynamically creates a Spark cluster on Aristotle cloud
  - Spark: A distributed compute engine for big data
- Cluster analyzes climate data available as distributed NetCDF files stored in an optimized format
- Sends results back to webGlobe client
Cloud Based Architecture

webGlobe Client

Data

User points to data

webGlobe Server

Access Data

Send data to client

Analyze Data

Spark
webGlobe Server Stack on Aristotle Cloud

- Allows dynamical provisioning of a virtual Spark cluster for analytics and visualization
- Enables parallel access to underlying massive simulation outputs
  - A novel spatial partitioning strategy

Try here: http://bit.ly/1XpFFCy
• NetCDF, HDF, LIDAR, Images, ShapeFiles, KML
• Thredds, WMS, OpenDap, HTTP, SSH, Local

Access

Visualize

Extend

Analyze

• Multiple projections
• Rectilinear, Curvilinear, Non-uniform grids
• Complex visualizations: animations, vector data, multi-level data

• Add web services
• Add server side analytics
• Customizable through configuration files

• Server side analysis – webGlobe Server
• Coming soon – R interface
Following the Cloud Principle

• Take computing to where the data is
  o Frees client resources

• Allows a way to extend
  o More analytics, support for data formats, etc.

• Well-equipped to handle the big data challenge

• A promising gateway to data archives
  o Exposes data and computing to the research and general community
Acknowledgements

• Funding: NSF CC*DNI DIBBs: Data Analysis and Management Building Blocks for Multi-Campus Cyberinfrastructure through Cloud Federation.

• Collaborators:
  o David Lifka (Cornell), Thomas Furlani (UB), Patrick Hogan (NASA)