Geoanalytics - A Solution Driven CI Platform for Geospatially Engaged Science

Jeff Heard, John McGee, Brian Blanton
RENCI, University of North Carolina at Chapel Hill

EarthCube White Paper Submission; October 17, 2011

Abstract: This whitepaper describes a Geoanalytics Platform under development as an abstraction of CI solutions for a number of science and research efforts. Many science domains are increasingly dependent upon geospatial data which has become a Big Data challenge. By combining internet scale Free and Open Source Software (FOSS), open source GIS, OGC standards, and advanced client applications with a hosted environment and cyberinfrastructure engagement experts, we can rapidly prototype and deploy advanced analytics solutions utilizing disparate large scale geospatial data.

1. Introduction

The pressures of producing science with global relevance and global impact have made understanding and using geographic information essential to a large portion of scientific research. Geographic information systems live at the heart of projects in public health, environmental science, policy and government, situational awareness, and others. Datasets essential to these projects are often terabytes in size, or rapidly evolving streams of complex data. Geographic data for science has become a Big Data challenge [1].

The tools available to professionals looking to *do* things with geographic data have not grown to meet the Big Data problem. Traditional GIS software enables researchers to build custom databases with analytics, and internet mapping services such as Google Maps allows them to publish data to the web. Various open source tools exist for specialized and general GIS purposes. As a result, geographic solutions to scientific and social problems are often cobbled together, resulting in silos that cannot be easily integrated or adapted to new and different data, or emerging management and analysis paradigms such as NoSQL [2] and Hadoop.

Traditional GIS solutions like ArcGIS and GRASS allow a user to perform complex analysis; however the results are treated as ends unto themselves, and not as data and methodologies to use in other ways. Modern users expect integration and web-based application platforms that include bleeding edge data and techniques. Solutions, such as Google Maps and Google App Engine allow a user to quickly create a map without prior training, requiring only a text editor, a web browser, and some patience. These solutions, in their goal for simplicity, abstract away functionality that is needed for serious scientific analysis. Therefore analysis must be performed first with other tools and then imported into data formats tailored towards visual presentation such as KML. These largely do not preserve the data for analysis or input into other models, adding complexity to the scientific process as well as discouraging the sharing of source data.

2. Driving Applications

The Geoanalytics CI Platform is being used in some capacity in the following projects:

- NCB Prepared prototypes; http://ncb-prepared.org/
- WxEM project: collaborative agreement between RENCI and NOAA for emergency management and situational awareness.
- UNC School of Public Health global impact website: http://www.sph.unc.edu/research/where_we_work.html
- UNC Gillings School of Public Health Farmers' Market Locator
- NARA's collaborative agreement with RENCI for scaling archiving cyberinfrastructure to billions of archival records, named CI-BER.

Geoanalytics is well-suited to projects in the environmental sciences, public health, and other projects requiring scalable cyberinfrastructure involving geography. It has been used to house and provide online access to such diverse data as:

- Environmental modeling data, including SLOSH and ADCIRC
- Pointwise data for public health applications
- NOAA weather forecast data
- Census data
- Very large and diverse archival metadata collections
- DOT road maps
- A complete set of 20m/pixel resolution LiDAR data for the state of NC

3. Geoanalytics Architecture

The Geoanalytics CI Platform is a novel solution for working with geographic data to solve a broad array of science challenges.

3.1 Platform Goals

The goals for this platform include:

- Scale horizontally to Big Data, its update frequency, access patterns, and management
- Integrate sensible data management solutions to scale
- Vet and federate Free and Open Source Software (FOSS) tools to lower the barrier-of-entry to using big geographic data, leveraging academic and internet scale software and infrastructure solutions
- Provide pathways, best practices, and modules to accomplish common tasks
- Allow users to rapidly develop and deploy prototypes and finished solutions
- Maintain a large scale hosted solution available and open to academic researchers to drive new requirements and overcome the activation energy required for working with large systems, while also contributing the blueprint and software necessary for advanced groups to deploy their own instance and/or contribute to the platform

The result of building a platform to meet these goals is a software ecosystem providing modular services for manipulating geographic data. Broadly, the platform breaks down into four layers, shown in the Figure 1:

- Data management and analytics layer incorporating iRODS, open source GIS software, distributed task queue, local and national compute resources
- Distributed geographic data models that encompass common data patterns
- Rapid web-application development platform based on open source GIS tools including implementations of standards such as WFS, WMS, and WCS layered on top of the data models
- A set of recommended and pre-integrated client-side technologies that form a core client layer for rapidly developing browser or mobile web-based applications

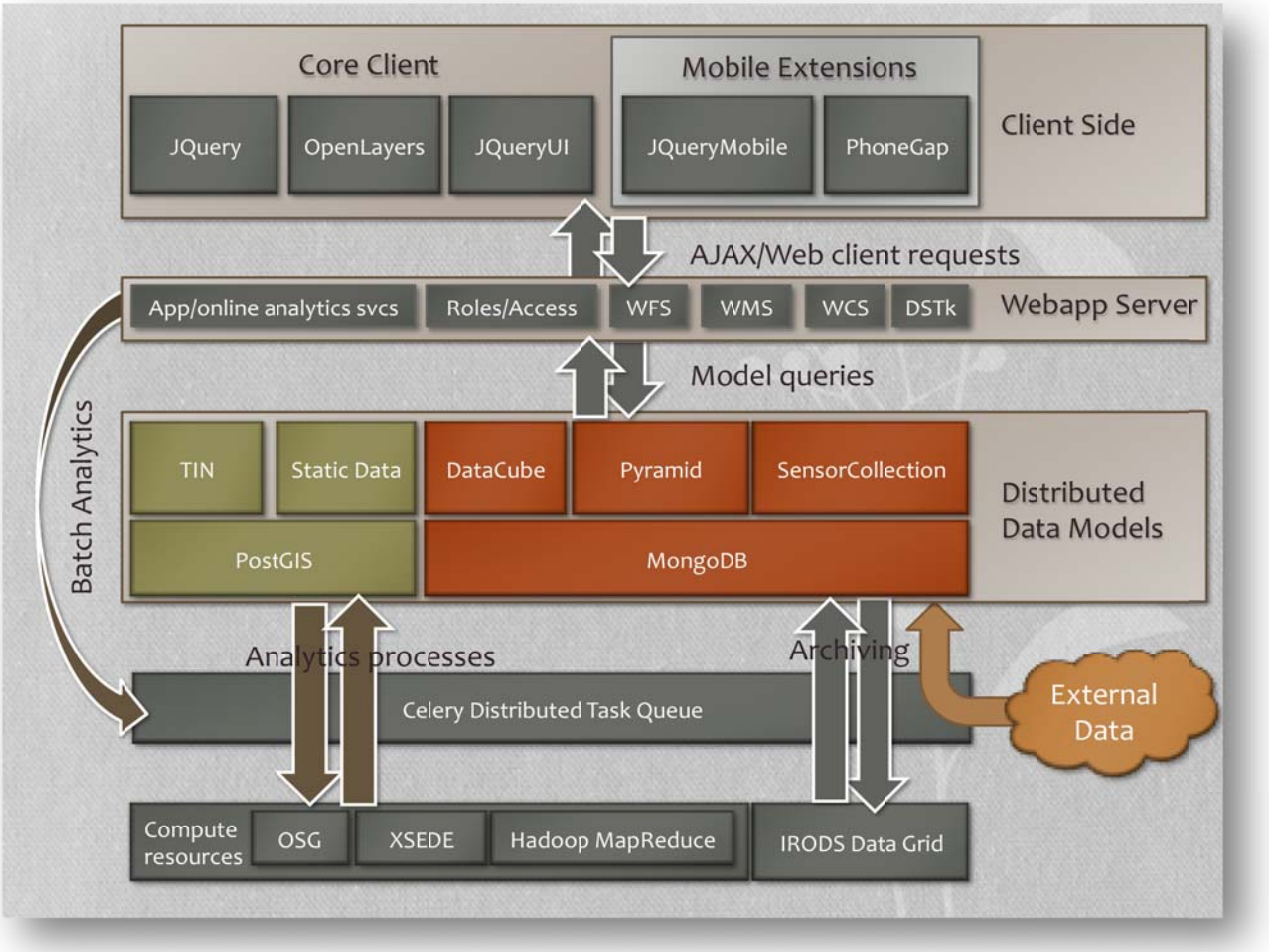

Figure 1 Architecture of Geoanalytics

3.2 Data management

The Integrated Rule-Oriented Data System (iRODS) provides a number of key capabilities such as encoding data management policies into rules that execute transparently as the iRODS virtual file system is used. This achieves two important goals: co-locating the logic of data processing with the data, encouraging data interoperability, and providing that policies are executed on all data in the system without human intervention. To move data to and from various iRODS Data Grids into and out of indexed, online random-access data models, we employ a distributed task-queue that uses rules to determine how and when to retrieve data. This task queue can be used for tasks such as refresh of data from remote sites, or sunsetting obsolete data by archiving or deleting it as determined by policy resolution.

3.3 Open data interchange and access standards

Critical to building large, enterprise or web-scale applications and community development around an open source platform is the ability to speak common data interchange languages. GIS has traditionally faltered in this area: geographic formats are almost as numerous as there are geographic projects; even single government agencies have been known produce data in multiple incompatible formats. As a result, the first step in many projects which include data from multiple domains is a data-import step where data sources are normalized into a single common format and projection scheme. Recently, the Open Geographic Consortium (OGC) has introduced a number of standards to provide for interchange of data, and from these a number of web services have become important to providing applications with data in the format they need. The Geoanalytics platform will support a subset of these standards that casts a wide net over common data access and interchange problems as driven by scientific use cases.

All models presented in the following sections support OGC's WMS (Web Mapping Service) for output, which provides styled visualization over web services. We extend WMS additionally to handle querying the underlying datasets and to handle in particular the time and elevation variables smoothly. Additionally, for TIN, SensorCollection, and custom models, we provide OGC's WFS (Web Feature Service) for accessing data its associated and geometry. Finally, for DataCube and Pyramid, we provide WCS functionality, which provides raster datasets containing underlying data values as opposed to the same data formatted for display.

3.4 Distributed Geographic Data Models

Geoanalytics provides a solution for managing and querying datasets on a distributed platform. Scalability is achieved using hybridized relational databases and non-relational data stores commonly known as NoSQL technologies, and Hadoop for very batch seek operations across very large data sets. There are naturally a number of common patterns in geographic datasets: Tiled data pyramids, common for large sets of high resolution satellite imagery; Spatial or spatiotemporal volumetric datasets, commonly used environmental, climatological, and geophysical datasets; The spatiotemporal feature collection, commonly found with sensor data; The feature collection, which contains static vector data; the coverage, which holds static raster data.

3.4.1 DataCube and TIN

Spatio-temporal volumetric datasets are common in environmental modeling and weather prediction. They are characterized by dense rasters of 2 or 3 dimensional raster data defined over a geographical area, and over a span of time. The DataCube model handles storage and selection of these rasters, providing for selection of data “swatches” across x, y, z, time, and version constraints. Swatches are returned as FORTRAN style arrays to the application programmer, or can be returned as NetCDF formatted data over the Web.

Another common environmental model is that of the Triangulated Irregular Networkor (TIN). A TIN is a densely packed but irregular raster where raster cells are of variable size of shape. This is very common for models with geospatially variable resolutions, such as ADCIRC and SLOSH. The TIN model handles this case, with an architecture and query mechanism similar to the DataCube, but paired with the interconnect network of X,Y,Z linkages between data points.

3.4.2 Pyramid

When dealing with high-resolution satellite or fly-over imagery, data stores designed for online access of imagery break it up into regularly sized tiles and then “mip-map” these tiles, creating a pyramid of tiles that can be accessed to stitch together imagery at different zoom levels quickly. The Pyramid data model is designed to handle just this case, where a large set of imagery, such as the orthophotography can be processed, hosted, and indexed, and accessed in a distributed manner.

3.4.3 SensorCollection

The SensorCollection model is a flexible data model designed to aggregate data from mobile and static sensor networks with static or ad-hoc membership. The SensorCollection can handle a traditional sensor network, such as an array of traffic cameras, *and* it can handle a rapidly evolving and changing mobile phone “crowdsourcing” network. SensorCollections also capture metadata and can store nested KVP metadata at the collection, sensor, or sensor-update level. Sensors send updates with arbitrary hierarchical or regular data, and the sensor collection can stream these updates in and index based on arbitrary data parameters as well as X, Y, Z, and time. Additionally, the SensorCollection model provides configurable strategies for reducing the “buildup” of sensor data in an online database. It provides services for archiving and sunsetting old data based on policies.

3.4.4 Custom PostGIS Models via GeoDjango

If none of the above cases will suffice, Geoanalytics provides the facility to create custom models that take advantage of all the other services Geoanalytics brings to bear. These custom models are indexed in a relational store or hybrid relational/non-relational store. As with all other models, custom data models can be used with Geoanalytics' built-in OGC compliant WMS and WFS services.

4. Communicating with Government and Business Decision Makers

Complex science and large volumes of disparate data can be very difficult to communicate effectively to government and business decision makers. An immersive visualization environment integrating data from high resolution imagery, sensed and measured data, model output, and more, that can scale from the desktop to large dome theater venues are a potential future method for greatly enhancing the impact and reach of these scientific endeavors. The WorldWide Telescope [3] from Microsoft Research is one example of an advanced client application that can consume data and services from the Geoanalytics Platform.

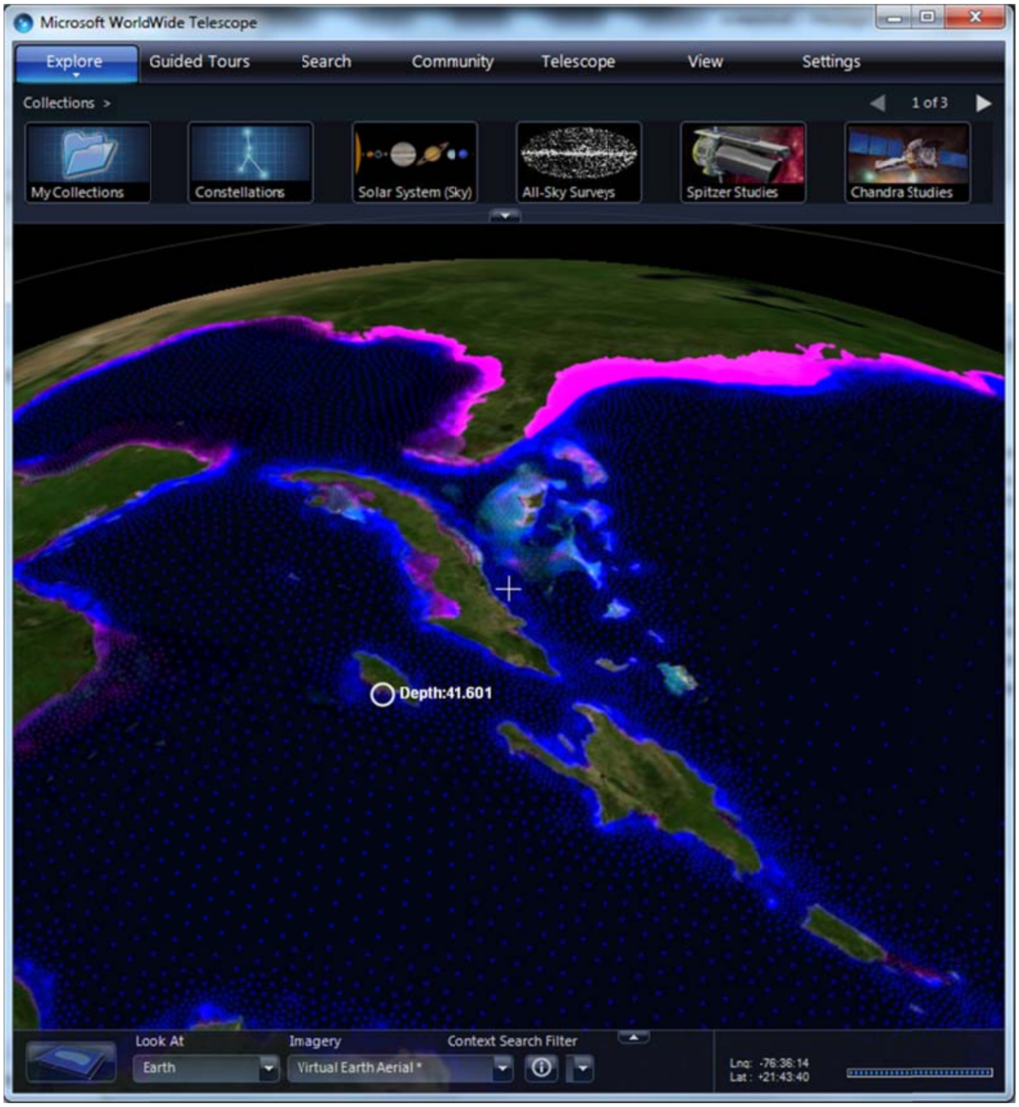

Figure 2: ADCIRC TIN file displayed in WorldWide Telescope (600k points)

Figure 2 above shows the ADCIRC TIN file that is used in numerous coastal science projects including FEMA floodplain mapping efforts. As demonstrated in Figure 3, this data can be overlaid with sensor point data, polygon data, and custom pyramid tile sets such as those processed and generated by the Geoanalytics Platform.

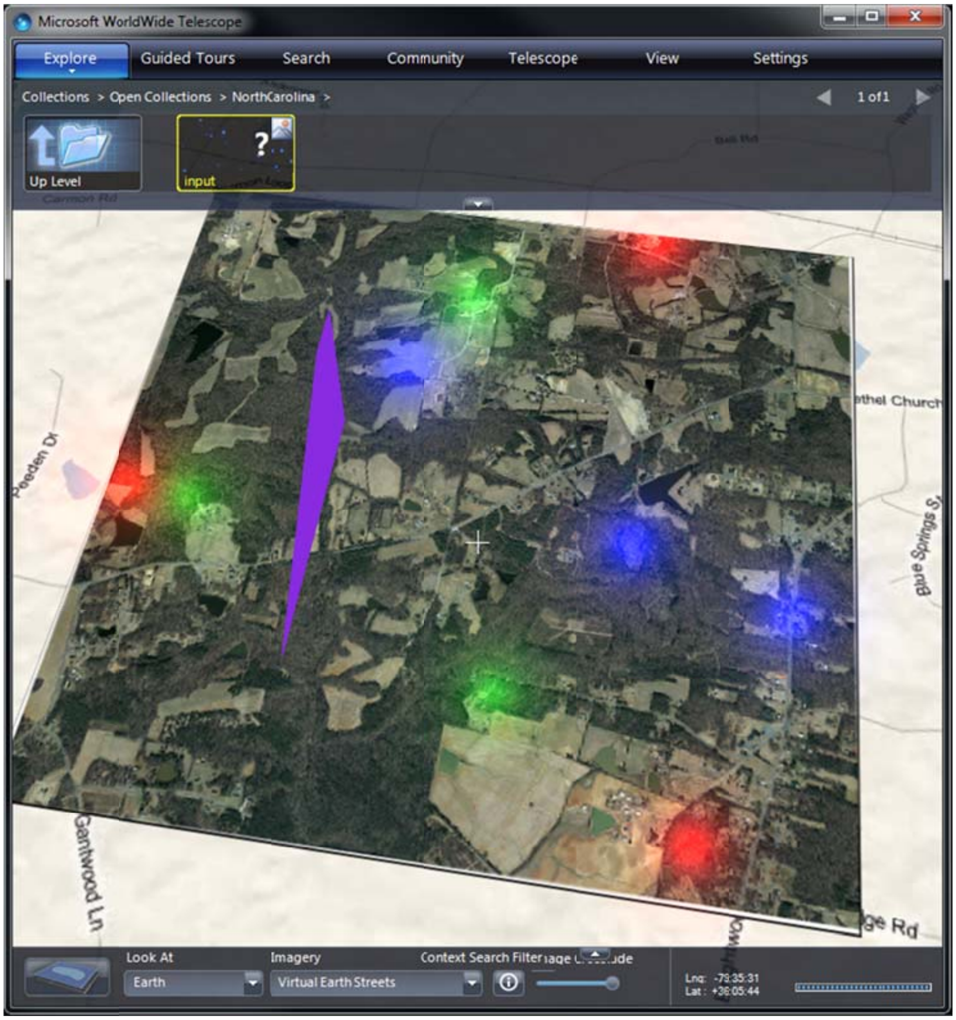

Figure 3. Point and area data combined with custom orthophoto overlay

References

[1] Lynch, C. Big data: How do your data grow? Nature 455 p28-29, 2008.
[2] Leavitt, N. Will NoSQL databases live up to their promise? IEEE Computer, 43(2), p12-14. 2010.
[3] http://www.worldwidetelescope.org/