GeoBOSS

Open-Source Geospatial Analytics at Scale

While geospatial libraries abound, no single library allows for classic geospatial coordinate transforms, geometric calculations, time-series analysis, machine learning methodologies, and ingest of open-source data from multiple sources without transferring between application programming interfaces and formats.

Through GeoBOSS, the Geospatial Analysis team at Pacific Northwest National Laboratory (PNNL) provides this interface to a growing number of analytics and data sources, including Global Administrative Boundaries, U.S. Census, Data.Gov, and others.

OVERVIEW

GeoBOSS is a software library that combines the data handling capabilities of Spark and the user-friendliness of Python to simplify geospatial analytics and easily transition between small-scale research and large-scale operational projects.

QUESTIONS

The initial use case for establishing the library focused on simplifying geospatial analysis that could scale to large data sets.
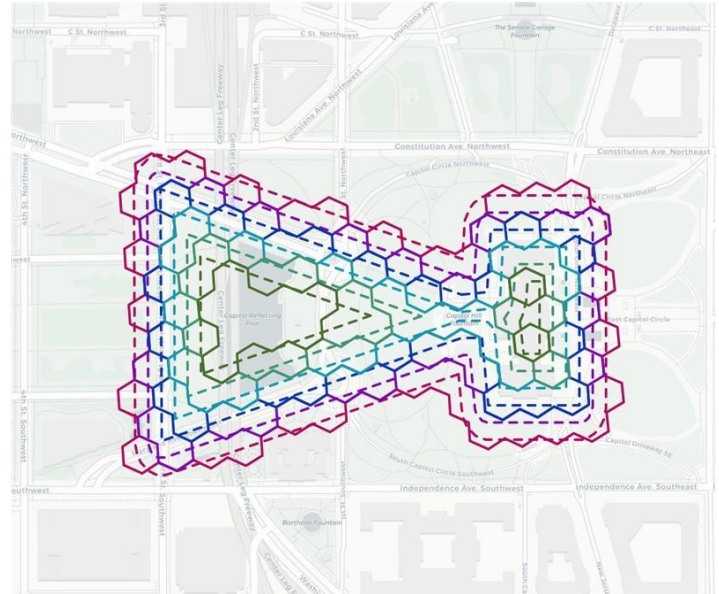

Today, PNNL's library helps analysts answer key questions:

- What infrastructure is impacted in a region hit by a disaster?

- Where do people travel and how will that be impacted by external events in an emergency?

- How do seasonal and daily trends affect use of physical infrastructure?

- Where is the greatest concentration of a needed service (like hospitals)?

OPERATIONS

PNNL developed GeoBOSS with the goal of making it broadly applicable to the geospatial community.
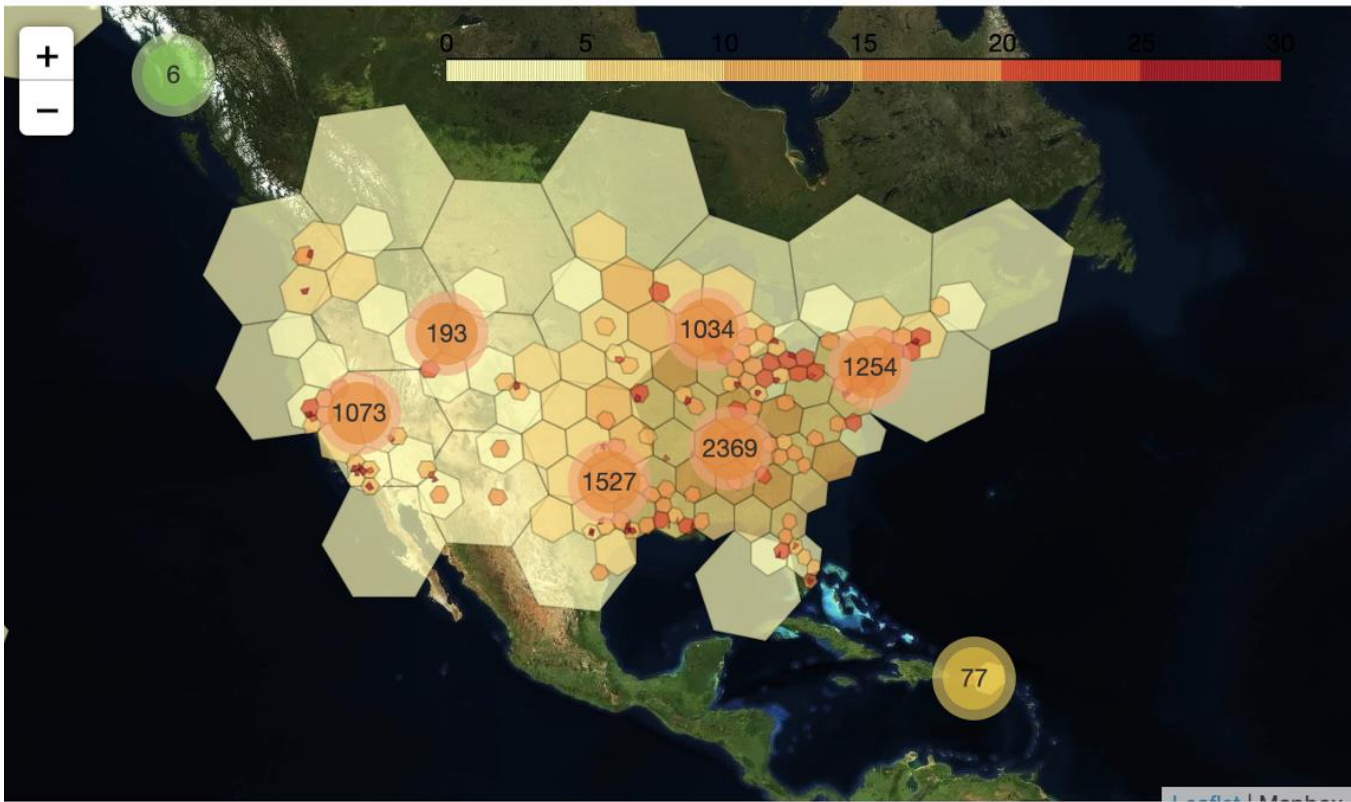
Within the library, individual data sources are transformed to abstract concepts of points, paths, and polygons. This helps users create analytics on a common framework that can be applied to many datasets and easily extended by the user. These range from simple feature generation (speed, bearing, etc.) to more advanced analytics (e.g., group movement identification, clustering, and other functions). Functionality is also included to simplify data clean-up, hashing, and plotting.

The biggest benefit of this library is its design to work with data at multiple scales—from small problems on a laptop to large database queries on a cluster—and in multiple environments, including local, cloud, and Databricks. This unified interface at multiple scales simplifies algorithm development and testing across datasets and environments.

BENEFITS AND IMPACTS

GeoBOSS provides the necessary flexible architecture to minimize data preparation for analytic development and mission needs.

PNNL's approach offers:

- Data ingest and standardization

- Fast searching and sub-setting

- Separate analytic development from specific data/use

- Scaling up data from bench scale to real data sizes

The library features an automated build and deployment pipeline, with pipelines made of building blocks that can be easily swapped out. The library also works within Databricks, local, or cloud environments.

For more information, contact:

Jenny Webster and Dan Corbiani

Computing and Analytics Division

National Security Directorate

geoboss@pnnl.gov

www.pnnl.gov