NGA Disparate Data Challenge Report

CyberGIS-Fusion

Our solution named as CyberGIS-Fusion is built as an app on the CyberGIS Gateway -- the leading online cyberGIS environment for a large number of users to perform computing- and data-intensive, collaborative geospatial problem-solving enabled by advanced cyberinfrastructure and big data capabilities. The CyberGIS Gateway is developed by our team, and supports access to a variety of geospatial big data such as agriculture, climate, Twitter, LiDAR, and multiple layers of the USGS National Map (e.g. 3DEP). The CyberGIS-Fusion solution takes advantage of the scalable data analysis, management, retrieval, and visualization capabilities of the CyberGIS Gateway with a particular focus placed on the selected datasets tailored to our specific answers to the challenge.

Our solution is focused on providing a scalable cyberGIS framework for flexible retrieval of disparate geospatial datasets while achieving near real-time query response and visual-statistical analysis. It is well known that all geospatial datasets can be modeled based on raster and vector representations. Our solution supports both data models in an integrated fashion. To demonstrate our unique capability of integrating and analyzing disparate datasets, we choose the city of Beijing as the area of interest.

Data Type Overview

Raster	Vector
• WMS • Tile Mapping Service • GeoTIFF	• Shapefile • KML • CSV • Json/GeoJson

Selected representative datasets:

- Beijing taxi trajectories (CSV, served as both GeoJson and Tile Mapping Service)
- Beijing housing prices (Shapefile, served as Tile Mapping Service)
- Beijing bus stop locations (Shapefile, served as Tile Mapping Service)
- Global Human Influence Index (WMS)

Vector support

Our solution establishes a novel and scalable scheme to store and retrieve spatiotemporal vector data. It incorporates a geohash-based indexing scheme to efficiently organize arbitrary spatial objects. The data storage and retrieval are implemented as an event-driven service to support high-performance access to the data. The backend database is Apache Cassandra tailored to store spatiotemporal data. We take advantage of the wide row capability in Cassandra to store spatial objects close to each other, hence improving querying performance. One major strength of our vector processing capability is its support for handling datasets with different shapes and precisions, while providing fast querying capability for large datasets.

For this specific application, our solution performs fast spatiotemporal queries on the taxi data. Many users can simultaneously access the application, and each may chose a bounding box and select an hour and will be provided with the heatmap of taxi locations based on the user-specified spatiotemporal constraint. The application also integrates a series of analytics to give users a high-level understanding of the urban dynamics of user-selected regions (e.g. volume of measurements by hour of the day and by day of the week).

In addition, users are able to efficiently track moving objects. A user may select a timeframe (up to 6 hours) and track taxi locations. This capability allows users to gain insights into intra-city dynamics by understanding how different city locations are connected based on the taxi traffic.

Raster support

Our solution provides a layer-based mapping service platform where users can add customized raster layers from various sources, including WMS, raw raster data, and analytical results derived from vector data. Users can also manipulate layer visibility within the platform to retrieve information by comparison.

Tile Mapping

Tile mapping services are mainly used to provide efficient map rendering of analytical results. A tile mapping server is deployed on ROGER (the first cyberGIS supercomputer), and is used to support mapping services for the CyberGIS-Fusion app. In our backend, analytical results (raster) are first generated as GeoTIFF. Then they are tiled on our mapping server and further published through web services.

Taxi Density

This capability demonstrates the spatial density maps of taxis. A user can query based on time on our app web interface to find the required density maps. Specifically, the location of each taxi within an hour is first extracted from the Beijing taxi trajectory dataset. In order to do so, a 10-minute moving window is set. If a taxi has more than one record within the time window, its location within this hour is defined as the location of records whose time is nearest to the hour. After that, a spatial density map of all taxi locations is generated through kernel estimation approach for each hour.

Taxi Locations

In this analysis, the locations of all taxis in each one-hour time span are displayed as a dot map. A user can query based on time on the app interface to find the required dot maps. In each one-hour time span, a dot map is visualized as a GeoTIFF image based on the locations of all taxi records during this time period.

Housing Price Surface

The housing price surface map shows the spatial trends of housing price (CNY/m^2). This trend surface is estimated using data 8 in the released dataset from the Beijing City Lab. The original data include the lat-lon of each location and the housing price. A spatial surface is then estimated using these point records through spatial smoothing and visualized on the app.

Bus Stop Density

The bus stop density map shows the spatial distribution of bus stops in a city. It is estimated using data 18 in the released dataset from the Beijing City Lab. The original data include the location of each bus stop. Spatial density maps are created using kernel density estimation.

WMS Services

In this prototype, we have a raster query service implemented on an external WMS service layer -- The Global Human Influence Index Dataset. It is a global dataset of 1-kilometer grid cells, representing the level of human influence on the environment. The value is calculated based on multiple data sources (e.g., land use, population density) and normalized between the range from 0 to 100. Higher values represent more human influence. The raster query is implemented so users can have a pie chart view of different human influenced regions within a selected

bounding box. We also implement a simple WMS import function so that users can load and visualize WMS services from external resources.