Geospatial Analytics in the US Environmental Protection Agency

Qi Dai
Senior Consultant, CGI Federal
Technical Lead, National Geospatial Support (NGS) Team
National Computer Center (NCC)
Office of Environmental Protection (OEI)
US Environmental Protection Agency

Presentation at the Center for Geospatial Analytics, College of Natural Resources, NCSU

March 26, 2015
Outline

• EPA National Computer Center and other EPA Organization Structures
• Geospatial Analytics Program in the EPA
• Examples of EPA Geospatial Analytics Projects
• Future Directions of EPA Geospatial Analytics Programs
• Questions and Answers
EPA Organizations

• EPA Mission
  – The mission of EPA is to protect human health and the environment

• EPA Organizations
  – 12 program offices
  – 10 regional offices
  – 7 research centers

• EPA Administrator: Gina McCarthy
  – was confirmed to her position in July 2013

• EPA employs roughly 15,000 full-time employees.
EPA 10 Regional Offices
EPA National Geospatial Support Team

- 15 members
- Design, operate, and maintain the GIS environment in the EPA NCC.
- Primary support for the EPA GeoPlatform System
- Manage Esri Enterprise Licenses for all EPA users
- Develop GIS apps for various EPA customers
- Develop and maintain EPA Shared Enterprise Geodata and Services
- Manage EPA national geospatial help desk
Geospatial Analytics Programs in the EPA

• EPA Office of Environmental Information (OEI) started the EPA Advanced Data Analytics and Visualization (ADAV) Program
  – Set up the ADAV development and test system in the EPA NCC
  – Conduct initial development projects to test various geospatial analytics technologies
  – Help other EPA program offices, regions, and research labs to utilize new geospatial analytics technologies.

• EPA Office of Enforcement and Compliance Assurance (OECA) utilizes geospatial analytics technologies for its Intelligent Targeted Enforcements
  – Use OSHA data to target RCRA inspections
  – Detect fraud in CWA discharge monitoring reports
  – Correlate water quality measurements with water quality criteria for better targeting

• EPA Office of Transportation and Air Quality (OTAQ) utilizes geospatial analytics technology to analyze ship AIS data for better air emission modelling on the coasts

• Data analytics and visualizations tools from EPA GeoPlatform
• The EPA Advanced Data Analytics and Visualization (ADAV) development and test system in the EPA NCC
  – 24 CPU cores, 52 GB memory, many TB storages, more resources can be added easily when needed.
  – Hortonworks HDP 2.2; Spark 1.2.0; Tableau 8.3; ArcGIS 10.3, etc.
  – Conduct initial development projects to test various geospatial analytics technologies
  – Help other EPA program offices, regions, and research labs to utilize new geospatial analytics technologies.

• Advanced Data Analytics and Visualization pilot project with EPA Air Quality Monitoring data
  – Apply an anomaly detection algorithm on spatio-temporal
  – Data is collected hourly by thousands of monitors and
  – Hadoop & Spark for batch analysis on large volume of data
  – Visualization of temporal results in Tableau and mapping
  – Export anomaly datasets to ArcGIS & GeoPlatform Online
  – Hortonworks Data Platform (HDP) cluster hosted in the NCC
EPA GeoPlatform

**EPA GeoPlatform Hosting Environment**
- **Internet** production servers
  - http://gispub.epa.gov
- **Intranet** staging and production servers
  - http://gisint#.rtpnc.epa.gov
  - http://gisstg#.rtpnc.epa.gov
- **Intranet** sandbox servers
  - FME Desktop/Workbench
  - http://sandbox#.rtpnc.epa.gov

**Data Extraction and Translation Tools**
- **EPA Clip N Ship**
  - Internet
  - Your Data
  - Regional Data
  - Program Data
  - http://edg.epa.gov/clipship
- **EPA FME Server**
  - Intranet
  - Transform
  - Exchange Load
  - Aggregate
  - Distribute

**EPA Data Download Site**
- **EPA Public Downloads**
  - Internet
  - gdb, shp, kml, csv
  - http://edg.epa.gov/data/Public
- **EPA Restricted Downloads**
  - Authenticated
  - Requires user login
  - gdb, shp, kml, csv
  - http://edg.epa.gov/data/Limited

**Data Visualization and Analysis Tools**
- **EPA GeoPlatform Online**
  - Internet
  - Load shapefile, csv, txt, or KML files from your desktop for display and sharing within a web map
  - https://epa.maps.arcgis.com

---

esri

EPA GeoPlatform Online:
Examples of EPA Geospatial Analytics Projects

• Advanced Data Analytics and Visualization pilot project with the EPA Air Quality Monitoring data
• EPA EnviroAtlas Applications
• EPA CleanPowerPlan Mapping Application
• EPA Ground-Level Ozone Mapping Application
• EPA GeoPlatform OpenGIS Demonstration projects
• Other EPA Geospatial Analytics Applications
ADAVID Pilot Project with EPA AQS Data

AQS Pilot Project Workflow Overview

Data Producers: Air Quality Sensors

Collection: AQS Data Mart

Storage: Hadoop File System

Analysis: Spark, Hadoop, Stats/ML

Presentation: Graphics (Tableau), Maps
Detecting Anomalies in Historical Air Quality Monitoring Data – EPA OAQPS

- Analyze hourly data over 10 years for anomalies in 6 pollutants.
- Use Esri’s GIS Tools for Hadoop for spatial proximity analysis on NEI Facilities and NRC Events.
- Non-point sources such as road network spatial density and historic large wildfire locations in the future

Identify Anomalies

Locate Potential Cause

Tableau graphics for tabular data (charts, basic maps)

Map viewer with time slider and nearby facilities/events
**ADAV Pilot Project with EPA AQS Data**

**Esri Open Source GIS Tool for Hadoop**

```
<table>
<thead>
<tr>
<th>localStr</th>
<th>timeOfdayHHMM</th>
<th>qualifier</th>
<th>state_code</th>
<th>county_code</th>
<th>site_num</th>
<th>state_name</th>
<th>county_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>4:00</td>
<td>&lt;Null&gt;</td>
<td>48</td>
<td>61</td>
<td>6</td>
<td>Texas</td>
<td>Cameron</td>
</tr>
<tr>
<td>13</td>
<td>5:00</td>
<td>&lt;Null&gt;</td>
<td>48</td>
<td>61</td>
<td>6</td>
<td>Texas</td>
<td>Cameron</td>
</tr>
</tbody>
</table>
```
- Tableau analysis for anomalies per monitor by pollutant
- CO and SO2 have most anomalies
- One specific monitor (270530954) has most anomalies
- One specific monitor (300870762) has most NO2 anomalies
Under President Obama’s Climate Action Plan, the U.S. Environmental Protection Agency proposed a common sense plan to cut carbon pollution from power plants. The Clean Power Plan will maintain an affordable, reliable energy system, while cutting pollution and protecting our health and environment now and for future generations.

It was developed and deployed very quickly for a major EPA press release.

Received very high web traffic initially.

Great collaboration efforts between EPA OAQPS staff and EPA NGS Team.
EPA Ground-level Ozone Mapping Apps

- Developed and deployed very quickly to support EPA press release
- Great collaboration efforts between EPA OAQPS staff and EPA NGS Team
- Interactive slider mapping apps allows users to compare two maps
  - before and current
  - current and future
EPA OpenGIS Demonstration Projects

- Download Boundless OpenGeo from the boundlessgeo.com website.
- Free and require no license fees.
- Installation of the Boundless OpenGeo Suite is easy and simple.
- It is easy to start the GeoServer with the default setting. Customized configurations and managements require a lot learning and trainings.
The Environmental Dataset Gateway (EDG) in the EPA NCC has been using the PostgreSQL database for many years.

The EPA National Geospatial Support (NGS) team completed the Standard Configuration Document (SCD) for the PostgreSQL use in the EPA. This SCD has been approved by EPA OEI/OTOP as part of the process to add the PostgreSQL to the EPA IT Architecture Roadmap.

The EPA National Geospatial Support (NGS) team has installed and operate several PostgreSQL databases in the EPA NCC GIS development environments that can be accessed by EPA GIS users from program offices and regions.

EPA GIS users have the options to test, experiment, and learn the PostgreSQL databases in the EPA NCC GIS development environment before implementing it on their local GIS servers.

Administration of the PostgreSQL databases is much easier than other commercial databases such as Oracle. It is 100% open source and require no license fee.

Esri utilizes the PostgreSQL in many of its product development processes. The ArcGIS Suite works with the PostgreSQL very well.

The National Geospatial Support (NGS) team has completed the EPA NCC Standard Configuration Document (SCD) about PostgreSQL and have submitted it for approval and addition to the EPA IT Architecture Roadmap system.

The National Geospatial Support (NGS) team continues to provide technical support for the use of PostgreSQL to all EPA GIS users.
Comparison of GIS Services with ArcGIS Server and Open Source GeoServer

- 2 GIS services based on ArcGIS Server 10.22 and 2 GIS services based on Boundless GeoServer 4.1
- All 4 GIS services are using the identical data from PostgreSQL database.
- 20 simulated concurrent users during the test period.
- When the size of the service is at 3.6 kb, the GIS service from the GeoServer is about 50% faster than that from the ArcGIS Server.
- When the size of the service is at 9 kb, the GIS service from the GeoServer is about 35% faster than that from the ArcGIS Server.
- All 4 GIS services are small and simple. When the size of these GIS services increases, the difference between ArcGIS Server services and GeoServer services become narrowed.

<table>
<thead>
<tr>
<th>Name</th>
<th># requests</th>
<th># failures</th>
<th>Median response time</th>
<th>Average response time</th>
<th>Max response time</th>
<th>Average Content Size</th>
<th>Requests/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>/arcgis/rest/services/OSPilot/R1LocalTravelPG/MapServer/export?</td>
<td>135</td>
<td>0</td>
<td>3200</td>
<td>3885</td>
<td>11826</td>
<td>3628</td>
<td>1.03</td>
</tr>
<tr>
<td>/arcgis/rest/services/OSPilot/R1LocalTravelPG/MapServer/export?</td>
<td>102</td>
<td>0</td>
<td>3000</td>
<td>3648</td>
<td>9741</td>
<td>9724</td>
<td>0.78</td>
</tr>
<tr>
<td>/geoserver/postgis1workspace/wms?</td>
<td>155</td>
<td>0</td>
<td>1421</td>
<td>1593</td>
<td>7284</td>
<td>3544</td>
<td>1.43</td>
</tr>
<tr>
<td>/geoserver/postgis1workspace/wms?</td>
<td>135</td>
<td>0</td>
<td>1938</td>
<td>2041</td>
<td>6022</td>
<td>9288</td>
<td>1.25</td>
</tr>
</tbody>
</table>
EPA NEPAssist Apps

NEPAssist is a tool that facilitates the environmental review process and project planning in relation to environmental considerations. The web-based application draws environmental data dynamically from EPA’s Geographic Information System databases and web services and provides immediate screening of environmental assessment indicators for a user-defined area of interest.

EPA EJScreen Apps

EJSCREEN is an environmental justice (EJ) screening and mapping tool that provides EPA with a nationally consistent dataset and methodology for calculating “EJ indexes”.

Other EPA Geospatial Analytics Applications

EPA NEPAssist Apps

EPA EJScreen Apps
Other EPA Geospatial Analytics Applications

- MyEnvironment provides immediate access to a cross-section of environmental data for any geographical location in the U.S.
- Users can search by neighborhood and learn about the air quality, cancer risk levels, water quality, brownfields and other factors that determine the overall environmental quality and livability of a community.
Future Directions of Geospatial Analytics in the EPA

• Utilize external Cloud infrastructures for large scale EPA Advanced Data Analytics and Visualization projects.
• Continue to develop in-house geospatial analytics capacity.
• Provide EPA customized Advanced Data Analytics and Visualization training for many EPA users.
• Promote and encourage the development of EPA Advanced Data Analytics and Visualization projects for various mobile devices.
• Encourage more collaborations with other federal agencies, state partners, and academic communities.
Qi Dai
Phone: 919-294-3719
Email: dai.qi@epa.gov