GEOSPATIAL VISUAL ANALYTICS BELONGS TO DATABASE SYSTEMS

Jia Yu
Advisor: Mohamed Sarwat

Arizona State University
BIG SPATIAL DATA ERA

1.19 billion monthly active users as of September 30, 2013

Social Media

Scientific Data

Urban Data

Medical Data

Mobile Devices

400GB of Road Network Data

85 Billion neurons in the brain nervous system
NYC TAXI TRIPS

Over a billion taxi trips between 2009 and 2015 released by NYC taxi and limousine services

<table>
<thead>
<tr>
<th>Pick Up Time</th>
<th>Pick up Location</th>
<th>Drop Off Location</th>
<th>Fare amount</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Laguardia</td>
<td>Manhattan</td>
<td>$40</td>
<td>…</td>
</tr>
<tr>
<td>2</td>
<td>JFK</td>
<td>Empire State</td>
<td>$30</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Run a sequence of spatial database queries:
Return all the NYC taxi trips for which the pick up location is within the Laguardia airport region
HEAT MAP OF NYC TAXI TRIPS
GEOVIZ

Not Scalable

Google Maps

GeoSpatial Map Visualization Tool

Not Interactive

Spatial Database

GeoSpatial Data

Zoom-In

Pan

Zoom-Out
ON-TOP DB APPROACH

(1) Load / Parse data: HDFS, S3, PostgreSQL

(2) Spatial Query: PostgreSQL, Hadoop, Spark (GeoSpark)

(3) Then do data visualization using Google Map, MapBox, ArcGIS, MapD

Painful process of jumping between query processing and data visualization especially with Big Data
Spatial Datasets

Spatial Query Processing

Map Visualization

Data Prep Phase

Viz Phase

Scatter Plot  Heat Map  Choropleth Map  . . .

Point Data  Polygon Data  Rectangle Data  . . .
REGULAR EXECUTION PLAN (DISTRIBUTED)
Implement the Geospatial Visual Analytics functionality on-top of the database system
VISUAL ANALYTICS IN THE DATABASE

Reduce the overhead of loading the data to a Map Visualization tool
BABYLON

AN END-TO-END VISUAL ANALYTICS SYSTEMS FOR MASSIVE-SCALE GEOSPATIAL DATA
SELECT ScatterPlot_OSM_L6 (taxi.pickup) FROM NYCtaxi taxi
WHERE ST_WITHIN(taxi.pickup,ManhattanBound)
SELECT ScatterPlot OSM L6 (taxi.pickup) FROM NYCtaxi taxi WHERE ST_WITHIN(taxi.pickup, ManhattanBound)

| Pick Up Time | Pick up location | Drop Off Location | Fare amount | ...
|--------------|------------------|-------------------|-------------|-----|
| 1            | Laguardia        | Manhattan         | $40         | ...
| 2            | JFK              | Empire State      | $30         | ...
| ...          | ...              | ...               | ...         | ...|
The partitioning is balanced based upon both the Spatial proximity and the visualization constraints.
INTEGRATE OPERATORS
By adopting the GeoVizDB approach, Babylon can achieve more than an order magnitude faster Data to Visualization time than state-of-the-art GeoViz systems.
OPPORTUNITIES

• Leverage existing database ideas, e.g., materialized view maintenance to support dynamic GeoViz

• Leverage GPU-accelerated databases (e.g., MapD) for massive parallelization of data aggregation and map rendering

• More representative and accurate GeoViz-Aware Sampling Techniques

• Deduce Human-Map-Interaction prediction models
Babylon is open-sourced on GitHub
https://github.com/DataSystemsLab/Babylon

THE DATA SYSTEMS LAB

https://www.datasyslab.net
@DataSysLab