

Scalable, Server-side Mapping in Drupal with the Geocluster-Leaflet Stack

Eric Paul (@mpgeek)



Phase2 Technology

October 19, 2014



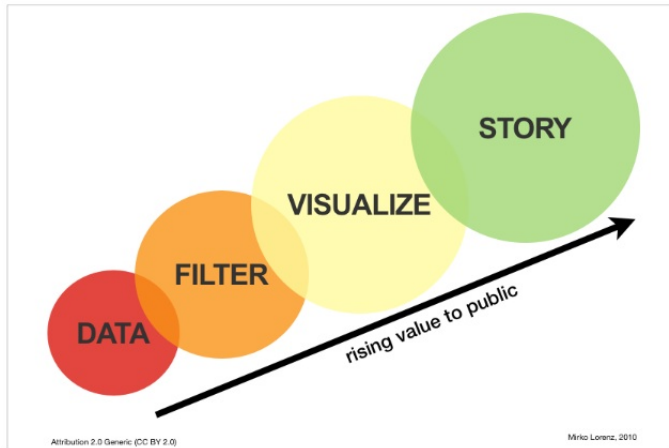
Table of contents

- 1 Background
- 2 The Geocluster-Leaflet Stack
- 3 Customization Towards an Application
- 4 Takeaways



Mapping: What Is Going On Here?

DATA-DRIVEN JOURNALISM = PROCESS



The Process

Data Driven Journalism as a **Process**

- Given a set of data
- And a question to be asked
- Find a useful visualization
- Data can then **tell the story**



The Process

Data Driven Journalism as a **Process**

- Given a set of data
- And a question to be asked
- Find a useful visualization
- Data can then **tell the story**

This allows content authors to **present data in context** in ways that would be **difficult with words alone**.

- We can adopt this media-industry notion and generalize it to **usability**.



Mapping: Why is This Important?



The Problem: Dense-Point Data

First pass: we have **point crowding**.

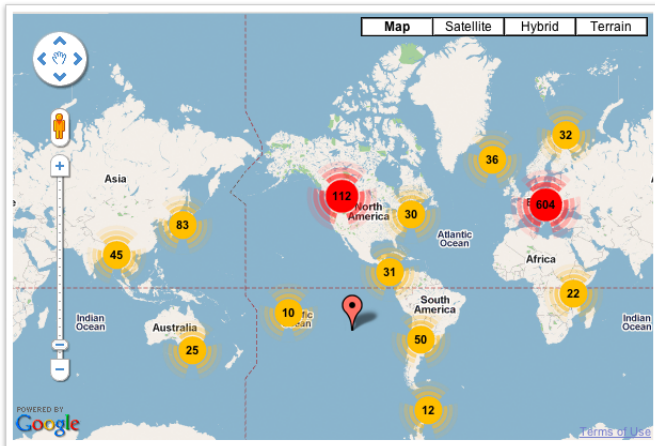


Really **not usable**.



One Solution: Client-Side Clustering

First step: lets cluster on the **client side**.

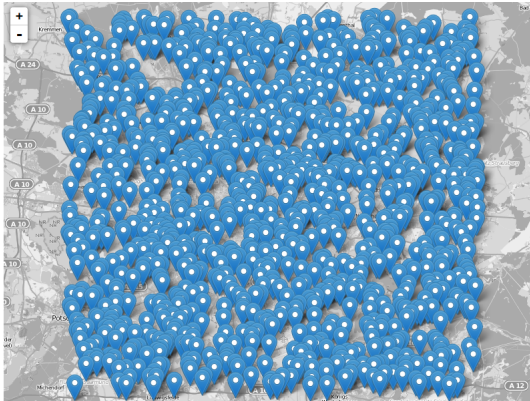


More usable, we gain context and can zoom in on areas of interest.



Solution Breakdown: Clustering Thousands of Points

What if we have **thousands** of points?



Client-side clustering **breaks down** upwards of a few hundred points.



Roadblock: Client-Side Clustering at Large Scale

Why Does it Break?

- 1 Views (**PHP**) renders each data point as a row of output, one at a time (thousands).
- 2 Views (**PHP**) renders the popup info (hidden) at page-load time.
- 3 The mapping library (**JS**) must parse the data.
- 4 The mapping library (**JS**) clusters the points.
- 5 The mapping library (**JS**) renders the map.



Roadblock: Client-Side Clustering at Large Scale

Why Does it Break?

- 1 Views (**PHP**) renders each data point as a row of output, one at a time (thousands).
- 2 Views (**PHP**) renders the popup info (hidden) at page-load time.
- 3 The mapping library (**JS**) must parse the data.
- 4 The mapping library (**JS**) clusters the points.
- 5 The mapping library (**JS**) renders the map.

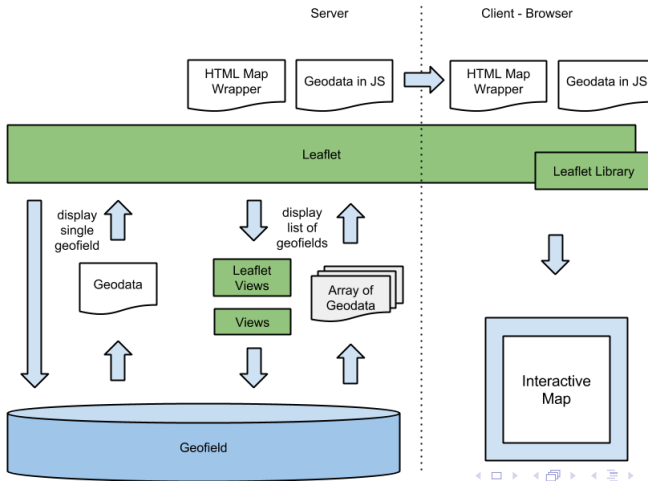
Both **PHP** and **JS** are asked to do too much at once.

- The **breaking point** is about 300 data points (empirical).



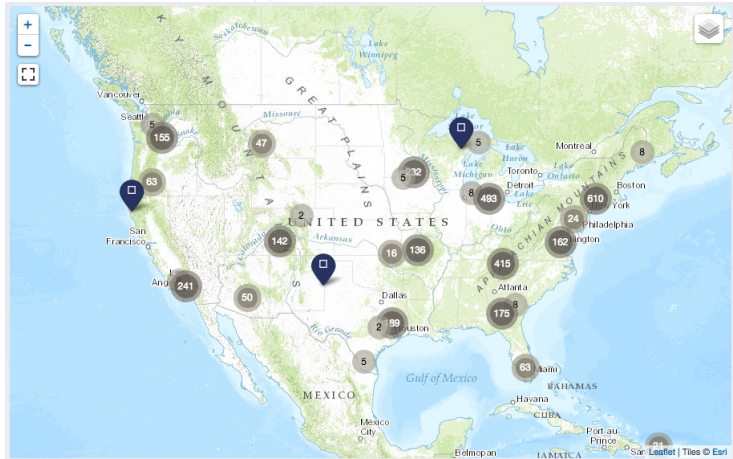
Client-Side Clustering Visualized

Drupal Mapping query and display modules - Leaflet



Demo

<http://vistacampus.gov/map>



Demo: Things of Note

- Bounded mapping (bbox strategy)
- Load time under 1sec
- Clusters are single things, not collections of things
- On-demand, ajax-delivered infobubbles
- Dynamic reclustering on pan/zoom
- About 4K points
- Layer interference (boo!)



Starter Build

If you really want to build this...

- 1 Clone the starter build
- 2 Modify to suit



Starter Build

If you really want to build this...

- 1 Clone the starter build
- 2 Modify to suit

Starter Build:

http://cgit.drupalcode.org/geocluster/tree/modules/geocluster_demo

- Instructions: <https://www.drupal.org/node/1962198>



Starter Build

If you really want to build this...

- 1 Clone the starter build
- 2 Modify to suit

Starter Build:

http://cgkit.drupalcode.org/geocluster/tree/modules/geocluster_demo

- Instructions: <https://www.drupal.org/node/1962198>

Why?

- The configuration is tedious and complex.
- Way too easy to break to start from scratch.



The Recipe

Basic Recipe

- Address Field (location storage)
- Geocoder (geocoding addresses, requires GeoPHP)
- Geofield (geocode storage)
- **Geocluster** (server-side clustering)
- Views
- Views GeoJSON (GeoJSON feeds)
- Leaflet GeoJSON (2.x for Panels support, 1.x for Bean)
- Leaflet Integration (requires Leaflet core library)



The Recipe

Basic Recipe

- Address Field (location storage)
- Geocoder (geocoding addresses, requires GeoPHP)
- Geofield (geocode storage)
- **Geocluster** (server-side clustering)
- Views
- Views GeoJSON (GeoJSON feeds)
- Leaflet GeoJSON (2.x for Panels support, 1.x for Bean)
- Leaflet Integration (requires Leaflet core library)

But... we need lots of **patches**.



A Working Model

The client build has been released as GPL2.0

- <https://github.com/mpgeek/VistaMap>

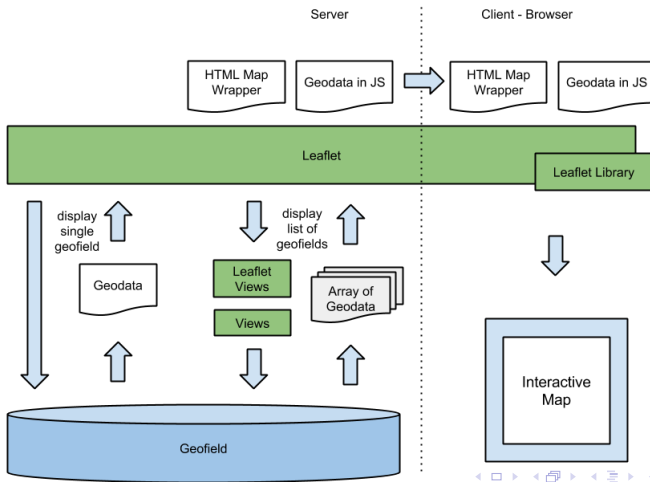
Patch mania! How about a **makefile**?

- https://github.com/mpgeek/Vista-Map/blob/master/vista_map.make

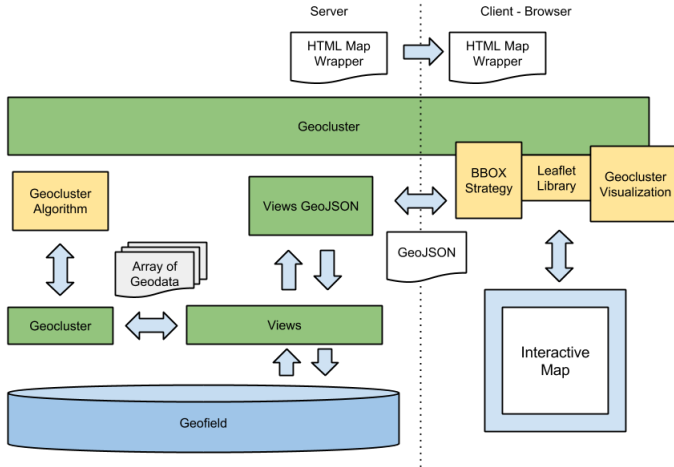


Client-Side Clustering Visualized (Redux)

Drupal Mapping query and display modules - Leaflet



Server-Side Clustering with Geocluster Visualized



Key Architectural Feature

Geocluster Keys

- Clustering is performed at the **query level** by Geocluster
- **PHP** and **JS** only see the clusters as single (Views) rows.
- This feature alone is almost **entirely responsible** for the performance gain.



Key Architectural Feature

Geocluster Keys

- Clustering is performed at the **query level** by Geocluster
- **PHP** and **JS** only see the clusters as single (Views) rows.
- This feature alone is almost **entirely responsible** for the performance gain.

But How?



Key Architectural Feature

Geocluster Keys

- Clustering is performed at the **query level** by Geocluster
- **PHP** and **JS** only see the clusters as single (Views) rows.
- This feature alone is almost **entirely responsible** for the performance gain.

But How?

- By **geohashing**!



Geocluster & Geohash

In a nutshell:

- Geocluster adds a hierarchical, spatial index to geofields based on the Geohash algorithm.
- Each geofield has columns for varying levels of precision (geohash index) created/updated on `entity_save`.
- A query for points/clusters specifies a geohash index and asks for clusters based on that index.



Geocluster & Geohash

In a nutshell:

- Geocluster adds a hierarchical, spatial index to geofields based on the Geohash algorithm.
- Each geofield has columns for varying levels of precision (geohash index) created/updated on `entity_save`.
- A query for points/clusters specifies a geohash index and asks for clusters based on that index.

Notice:

- The clustering information is created when the content is **created**.
- A request for points and clusters **doesn't actually cluster**. Rather it's a **simple query** of a spatial index.



Near-point Clusters vs. Exact-point Clusters

Monolithic Clusters

- Leaflet doesn't discern between points that are **near to one another** versus multiple points at the **same location**.
- We needed to create two cluster types, one for each condition.



Near-point Clusters vs. Exact-point Clusters

Monolithic Clusters

- Leaflet doesn't discern between points that are **near to one another** versus multiple points at the **same location**.
- We needed to create two cluster types, one for each condition.
- `vista_map.module`, lines 115-155



On-Demand Popups

AJAX!

- We don't load the popup info into the DOM at map-load time (performance tactic).
- We needed to load them on demand and allow them to be cached.



On-Demand Popups

AJAX!

- We don't load the popup info into the DOM at map-load time (performance tactic).
- We needed to load them on demand and allow them to be cached.
- `vista_map.js`, lines 324-404



Current-user Zoom

Focus the Map on the Current-user's Location

- One of the purposes of the map was to emphasize making local connections.
- We wanted to zoom in on the currently logged-in user.



Current-user Zoom

Focus the Map on the Current-user's Location

- One of the purposes of the map was to emphasize making local connections.
- We wanted to zoom in on the currently logged-in user.
- `vista_map.module`, lines 290-351



Limit Geocoder Granularity

Geocode to Center of ZIP-code Only

- One of two data layers needed to geocode only to ZIP-code precision.
- Removing more-specific information and passing abbreviated info only to geocoder.



Limit Geocoder Granularity

Geocode to Center of ZIP-code Only

- One of two data layers needed to geocode only to ZIP-code precision.
- Removing more-specific information and passing abbreviated info only to geocoder.
- `vista_map.module`, lines 12-72



Multiple Data Layers

Implement Data Layering and Panels Support

- OG membership drove layer membership, and source geofield.
- Views necessitated that different source geofields be separate data layers.



Multiple Data Layers

Implement Data Layering and Panels Support

- OG membership drove layer membership, and source geofield.
- Views necessitated that different source geofields be separate data layers.
- Contributed the 2.x branch of Leaflet GeoJSON for panels support with multiple data layers
(<https://www.drupal.org/node/2225815>)



Scalability Requirement

How big did we need to go?

- Mapping user profiles, about 18k users were migrated
- Originally, it was expected that all users would be map
- Application scale, then is 10^4



Scalability Requirement

How big did we need to go?

- Mapping user profiles, about 18k users were migrated
- Originally, it was expected that all users would be map
- Application scale, then is 10^4

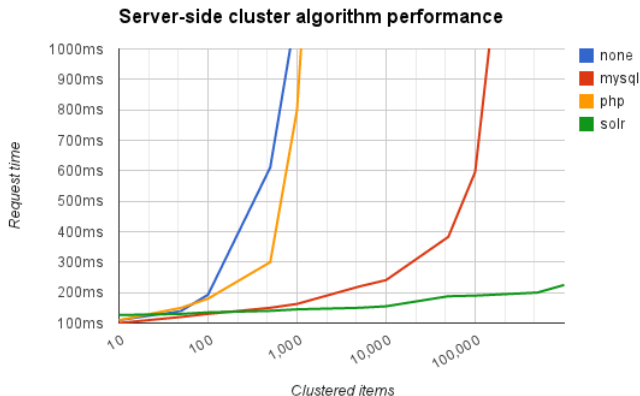
Geocluster's clustering backend is **pluggable**

- PHP clustering (post-query clusternig)
- MySQL clustering (query-level clustering)
- Apache solr clustering (alternative query-level clustering)



Scalability Metrics

Cold caches



We implemented **MySQL** clustering



Possible Improvements

Geocluster

- Progressively enhance with client side clustering below a certain point threshold.

<https://www.drupal.org/node/1914704>



Possible Improvements

Geocluster

- Progressively enhance with client side clustering below a certain point threshold.

<https://www.drupal.org/node/1914704>



Possible Improvements

Leaflet GeoJSON

- Collapse clusters to a single layer to eliminate layer interference.



Possible Improvements

Leaflet GeoJSON

- Collapse clusters to a single layer to eliminate layer interference.
- Make data feeds cacheable by quantizing bounding box parameters.

`/$view_url?bbox=$left,$right,$top,$bottom?zoom=$zoom_level`

- The `bbox` arguments are floating-point numbers that depend on viewport size and zoom. Takes a long time for caches to warm up for non-mobile viewports.
- <https://www.drupal.org/node/1868982>



Take-out Knowledge

What we know

- Large-scale mapping is now possible in Drupal.
- Geocluster needs work.
- Leaflet GeoJSON needs work.
- Despite that, production-quality map applications can now be built.



Take-out Knowledge

What we know

- Large-scale mapping is now possible in Drupal.
- Geocluster needs work.
- Leaflet GeoJSON needs work.
- Despite that, production-quality map applications can now be built.
- You will need a debugger.



References & Resources

Things we saw and more resources:

- Map application in production:
<http://www.vistacampus.gov/map>
- Map application Drupal feature:
<https://github.com/mpgeek/Vista-Map>
- Geohash Algorithm:
<http://en.wikipedia.org/wiki/Geohash>
- Geocluster Master's Thesis (by @dasjo):
<http://dasjo.at/thesis>



Questions?



Find me on twitter, IRC, or drupal.org: @mpgeek

