

Building an Open Source, Real-Time, Billion Object Spatio-Temporal Search Platform



2016 International Workshop on Cloud Computing and Big Data

Benjamin Lewis, David Strohschein, Paolo Corti, David Smiley
Center for Geographic Analysis, Harvard University

Background

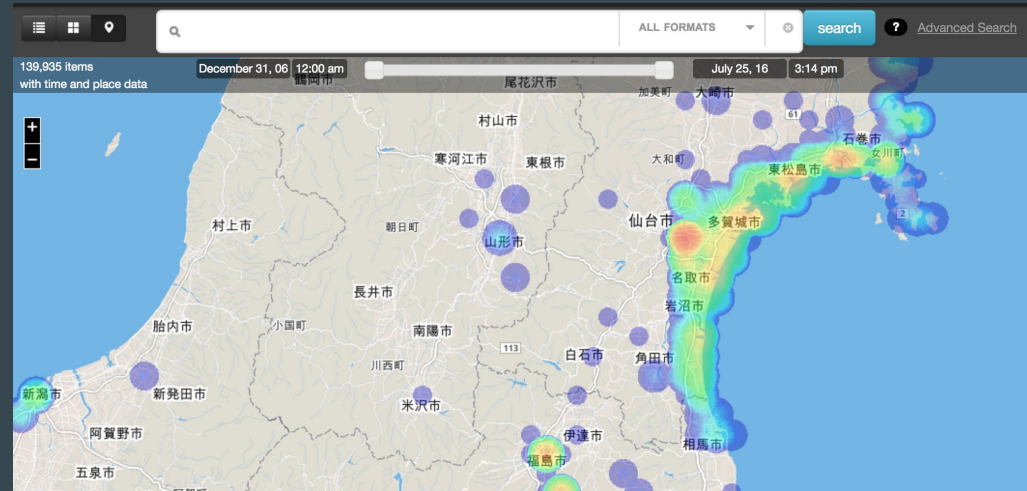
- Big data is everywhere: sensors (weather, pollution...), mobile devices, social platform activities, software logs, etc.
- Data are generally streaming, so they are temporal
- Most of those data are spatial as well
- Traditional RDBMS, desktop statistics and visualization packages have difficulty handling big data
- Current solutions involve “massive parallel software running on a large number of servers”

Use case

- We work in a research university so we need to provide big data to students and researchers
- Our goal is to lower barriers to interactive data exploration
- Some systems support visualization of large spatio-temporal datasets but don't handle search well
- Many search applications (most search engines) handle text but do not support the geographic dimension.
- Great need for tool to allow user to interactively search large collections and visualize them geographically. To support such increasingly common datasets, a new kind of map server and client is needed.
- Project funded by the Sloan Foundation in partnership with Dataverse team at Harvard IQSS

Solution

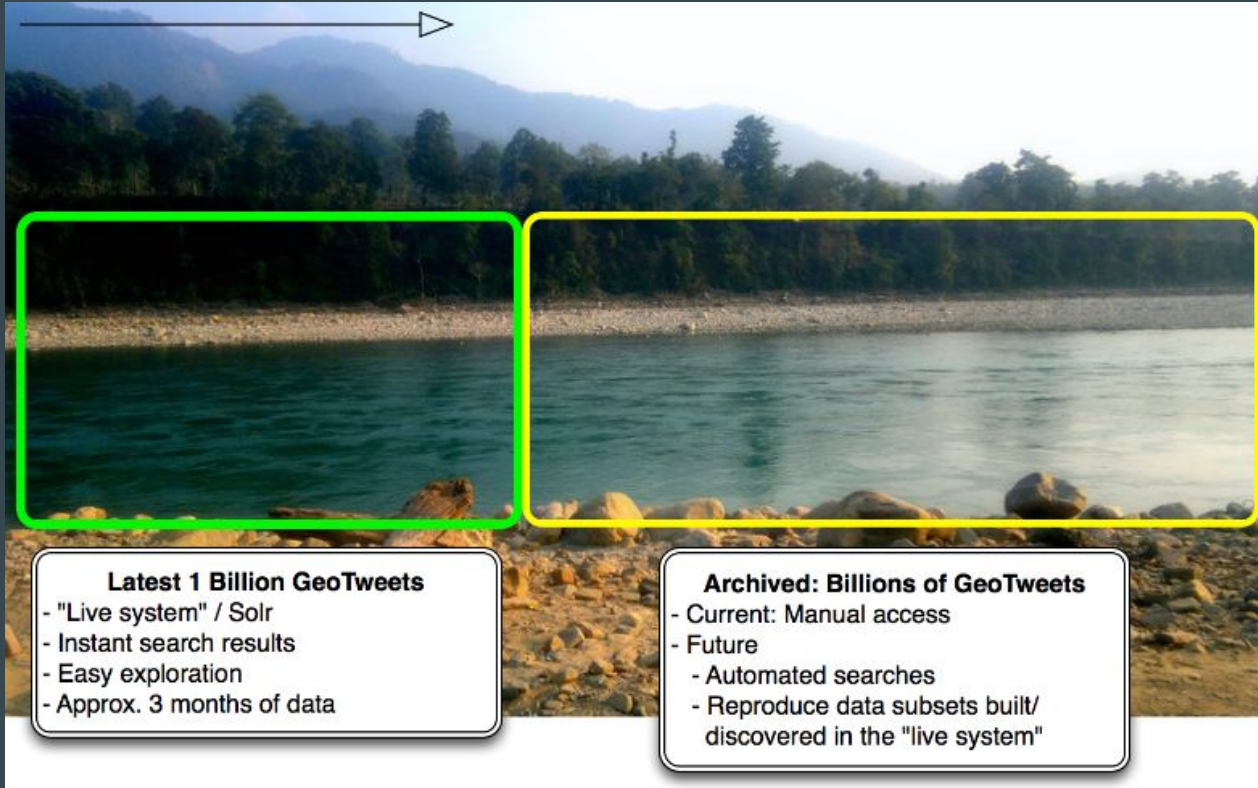
- A general solution. Prototype with geotagged tweets (tweets containing GPS coordinates from originating device)
- Platform adaptable to other big data spatial time streams (weather and pollution sensors, geoRSS feeds etc...)
- Integrate the new platform within Harvard WorldMap and Dataverse systems



Objective

- Create a missing piece of geo-infrastructure and make it available
- Demonstrate possibility of addressing scalability limitations with non-exotic software and hardware
- Make setting up platforms for big spatio-temporal visualization as easy as setting up a standard GIS stack

Streaming big data



←

Latest 1 Billion GeoTweets

- "Live system" / Solr
- Instant search results
- Easy exploration
- Approx. 3 months of data

Archived: Billions of GeoTweets

- Current: Manual access
- Future
 - Automated searches
 - Reproduce data subsets built/ discovered in the "live system"

Geotagged tweets

- Geotagged tweets: tweets containing GPS coordinates from originating device
- Currently about 2% of tweets are geotagged, about 8 million per day
- The CGA has been harvesting geo-tweets since October 2012 using the Twitter API
- Billion Object Platform(BOP) will provide a client and API to browse and search the latest 1 billion geotagged tweets (about 3 months range)
- Command line tools to extract older geotagged tweets from archives

The BOP (Billion Object Platform)

- General purpose, open source platform to support exploration of large collections of spatio-temporal entities
- Built on top of a search engine
- Supports exploration, visualization, extraction via a RESTful API
- Queryable by time, space, text
- Responsive
- Spatial heatmap to represent the distribution of results (spatial faceting: results per cell in a grid)
- Support temporal histograms (temporal faceting: results per date time range)
- Support word clouds as a mechanism to enhance results browsing by topic
- Support downloads of subsets for registered users (up to 10,000 features)
- Sentiment stamping

Solution Stack

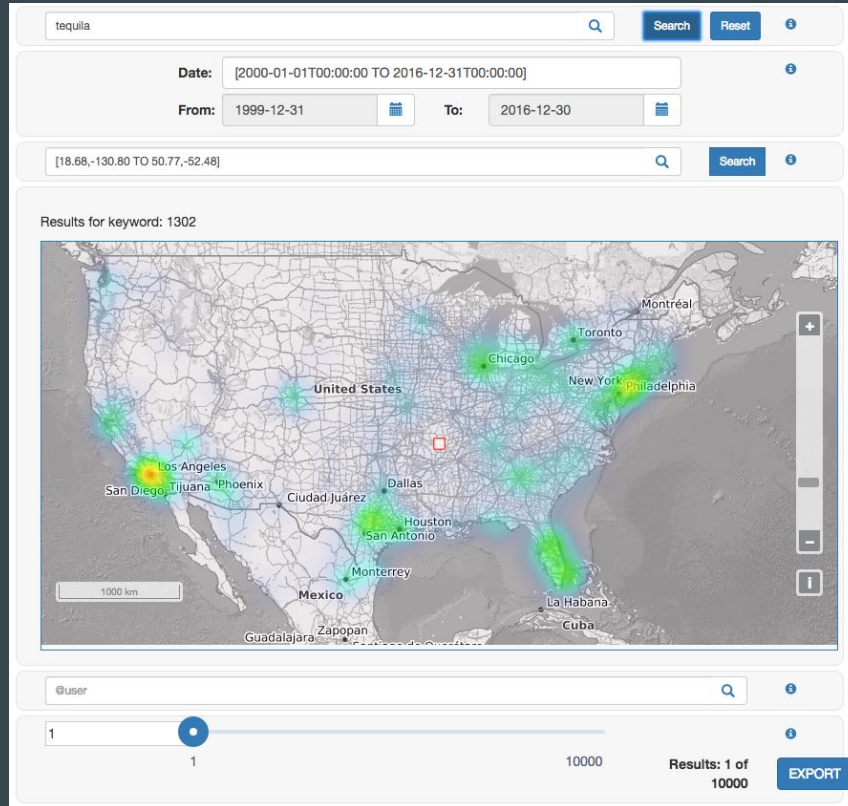
- Apache Lucene: an indexing and search library
- Apache Solr: a search web server platform built on top of Lucene
- Apache Kafka: a message broker written in Scala to provide a platform for handling real-time data streams
- Apache ZooKeeper: enables highly reliable distributed coordination
- Swagger: a framework for building APIs
- scikit-learn library: Machine Learning in Python
- OpenLayers: a javascript mapping client
- AngularJS: a javascript framework



Search engine features

- Faceted searches (category, space and time)
- Stemming: ability to detect words derived from a common root
- Synonyms detection and controlled vocabulary such as thesauri and taxonomies
- Weighted results
- Wildcard and fuzzy search: provide results for a given term and its common variations
- Boolean queries: search results using terms and boolean operators such as AND, OR, NOT...
- Hit highlighting: provides immediate suggestions to the user typing the text to search
- Stop words: words filtered out during the processing of text

Client to enable data exploration and extraction



API to streaming geotagged tweets

The image shows a Swagger UI interface for an API. At the top, there's a green header with the Swagger logo and the URL `http://54.158.101.33:8080/bopws/swagger.json`. Below the header, there are two input fields for API keys, both containing the text `api_key`, and an `Explore` button. The main content area is titled `default` and contains two API endpoints:

- `GET /tweets/export` with the description: "Search export endpoint for bulk doc retrieval."
- `GET /tweets/search` with the description: "Search/analytics endpoint; highly configurable. Not for bulk doc retrieval."

Below the endpoints, there are sections for `Implementation Notes`, `Response Class (Status 200)`, and `Model Schema`. The `Model Schema` section shows a JSON object:

```
{
  "a.matchDocs": 0,
  "d.docs": [
    {}
  ],
  "a.time": {
    "start": "string",
    "end": "string",
    "gap": "string",
    "counts": [
      ]
    }
  }
}
```

The `Response Content Type` is set to `application/json`. The `Parameters` section shows a table with one parameter:

Parameter	Value	Description	Parameter Type	Data Type
<code>q.time</code>	<input type="text" value="[2013-03-01 TO 2013-04-01T00:00:00]"/>	Constrains docs by time range. Either side can be '*' to signify open-ended. Otherwise it must be in either format as given in the example. UTC time zone is implied.	<code>query</code>	<code>string</code>

Sentiment Analysis

- Sentiment analysis is a field of study which identifies the opinion of people expressed in a text using natural language processing tools
- Social media such as Twitter provides a constant source of textual data, many with an opinion, which can be analyzed using Sentiment Analysis tools.
- Using the scikit-learn library (Machine Learning in Python) we sentiment stamp as positive or negative each tweet

Hypermap

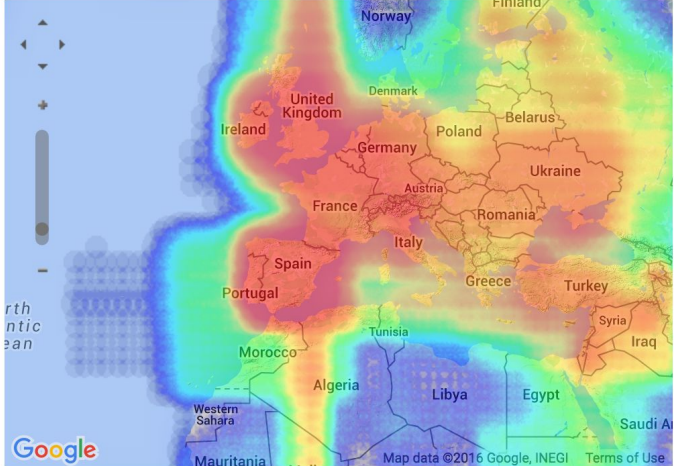
Similar approach to BOP (Solr/Lucene): provides a searchable registry of map service layers from OGC and Esri public endpoints

Search

SEARCH External Data Upload Layer Create Layer Rectify Layer

Keyword Source All Layers Search Reset

from year to year



Search results table:

Title	Source	Date
Distribution	warp.worldm...	None
Iron Age Cultures	worldmap.har...	None
Roman Empire: 117 CE	worldmap.har...	116
Roman Empire: 200 CE	worldmap.har...	199
Roman Roads	worldmap.har...	499
Roman Roads 2013	worldmap.har...	2012
Roman Empire: 69 CE	worldmap.har...	68
Roman Provinces ca. 303-...	worldmap.har...	300
Roman Empire: 14 CE	worldmap.har...	13
Roman Provinces (AD 100) Ba...	worldmap.har...	None
Roman Provinces	worldmap.har...	None
BA100 Provinces	worldmap.har...	None
Percorso	warp.worldm...	None
A new map of Turkey in Europe	maps.yelp.org	1699

Prev Next Showing 1-200 of 12747

No Layers Selected

Clear Selected Add To Map