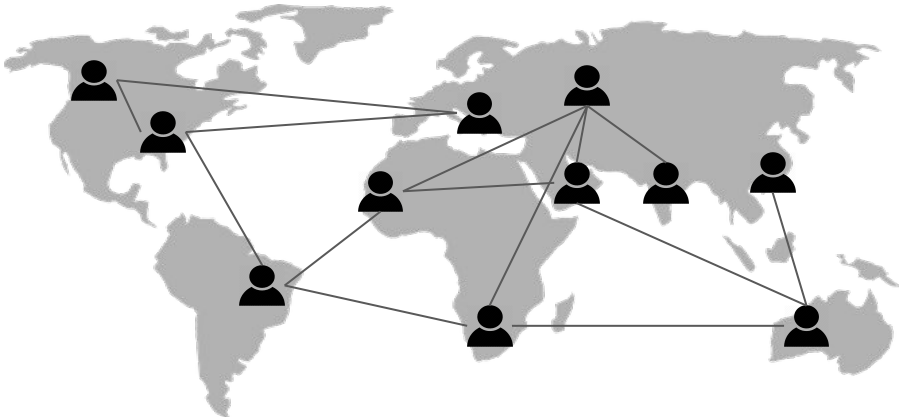
Spaten: a Spatio-Temporal and Textual Big Data Generator

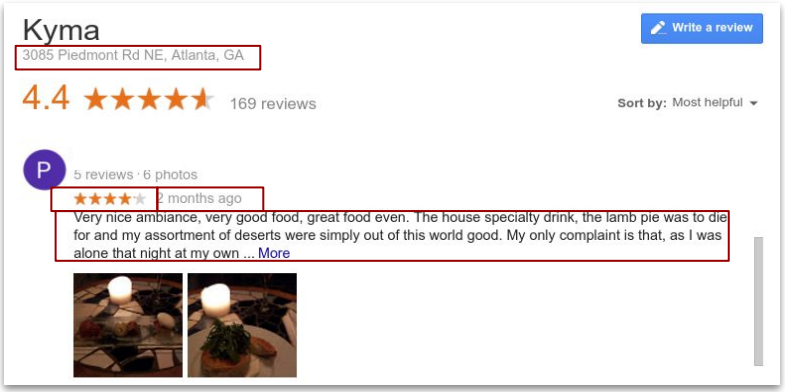
Thaleia Dimitra Doudali*
Ioannis Konstantinou
Nectarios Koziris

Motivation

1. Geo-Social Networking Graph

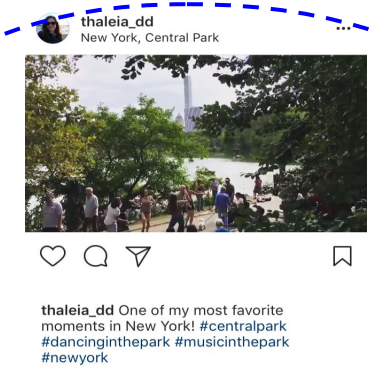


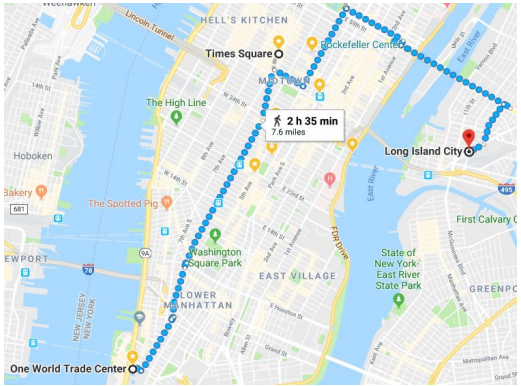
2. Spatio-temporal and textual data



Motivation

3. Daily routes with check-ins

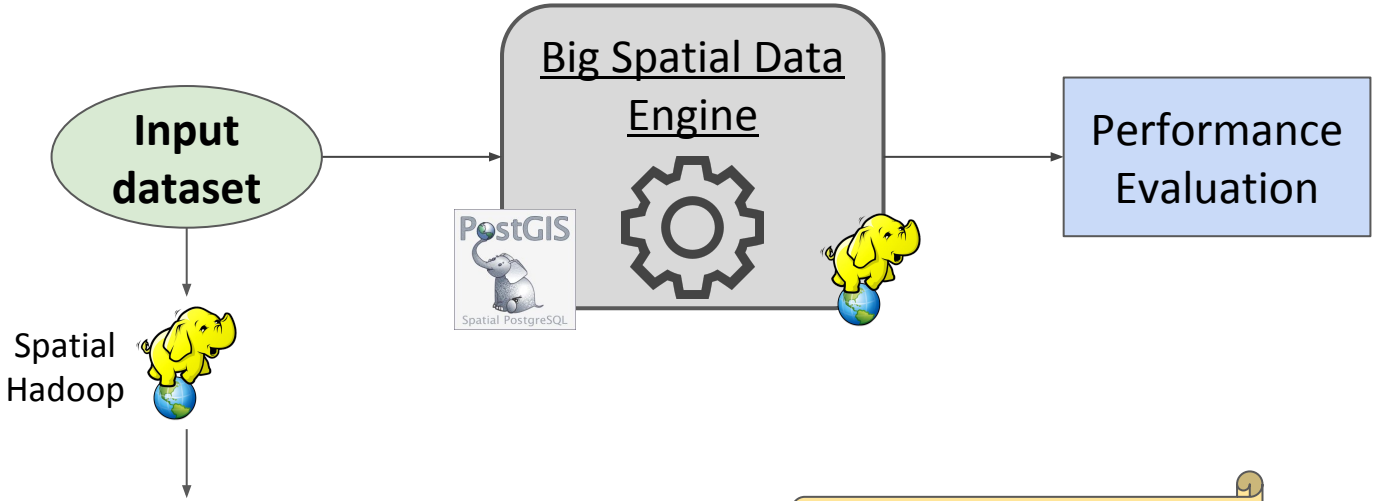




× millions of daily users
= part of Big Geo-Social Data

Motivation

New or extended Big Data Engines for *Spatial* data.

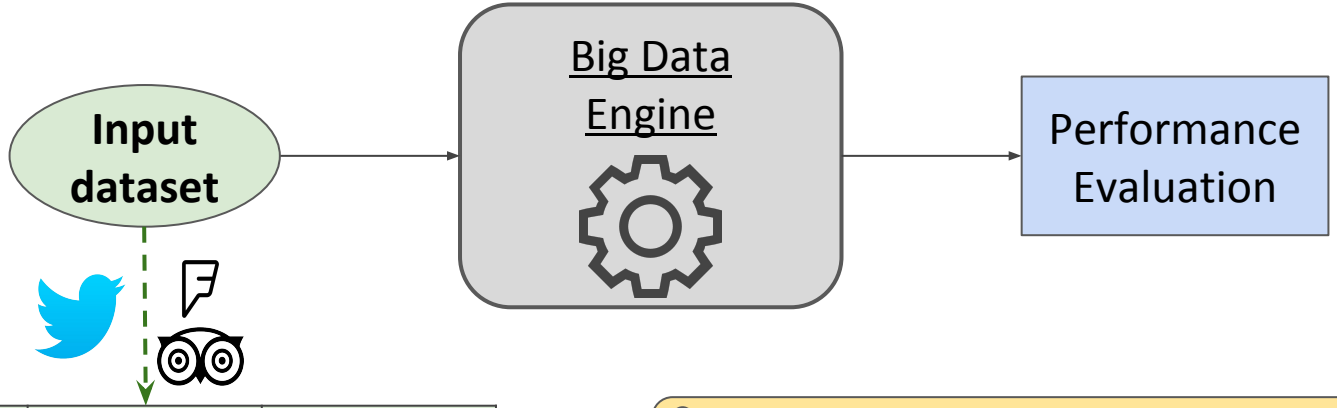


- OpenStreetMap (60 GB - real)
- NASA (4.6 TB - real)
- SYNTH (128 GB - synthetic)

Easy access to large spatial datasets. (real or synthetic)

Problem Statement

New or extended Big Data Engines for *Geo-Social* data.

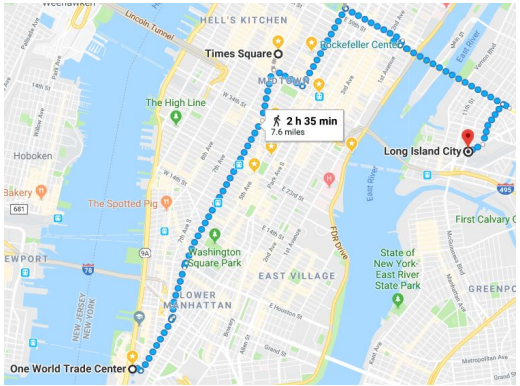


Type	Real	Synthetic
Small	✔	✔
Large	✖	✔

Can we create ***realistic*** (real source, synthetic combination) Geo-social data at a large scale, for performance and scalability evaluations?

Our Contributions

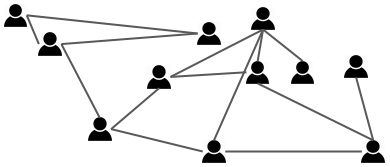
- Build ***Spaten***: a ***Spa***tio-***Te***mporal and Textual Big Data Ge***n***erator.
  - configurable, open source.



- Successfully create a large realistic Geo-social dataset.

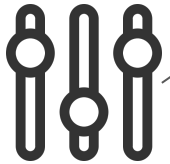
- Show how we can store and query the generated data, using state of the art NoSQL database systems.

Overview



1. Social network graph

2. Points of Interest (POIs)



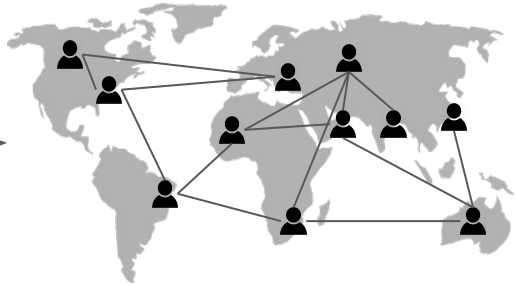
3. Configuration Parameters

Input

Spaten

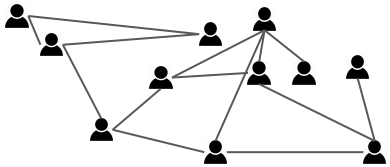
Creates daily routes with check-ins of users to POIs

Output

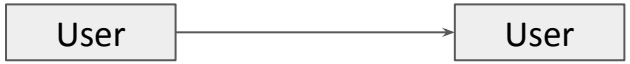


Geo-Social network

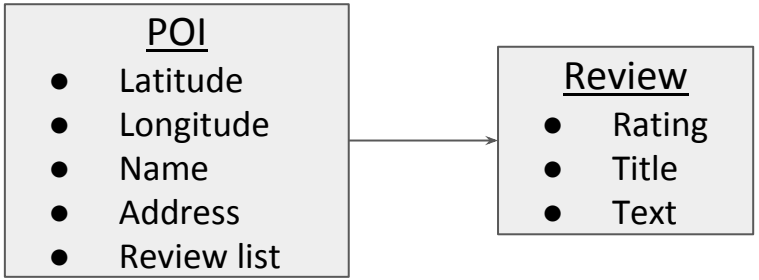
Input Data



1. Social network graph

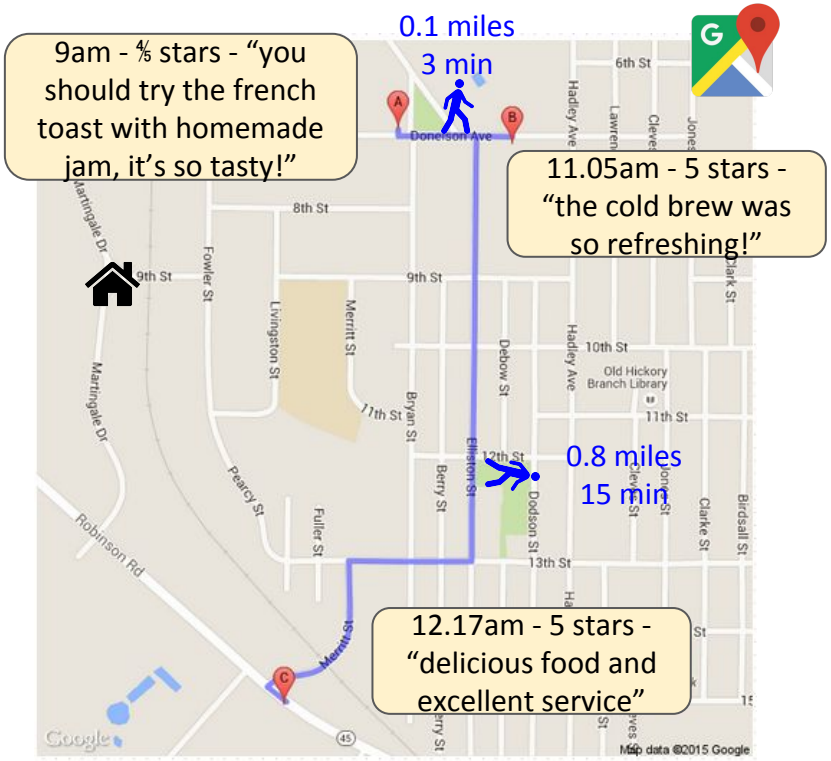


2. Points of Interest (POIs)



Data Generation Process - Example

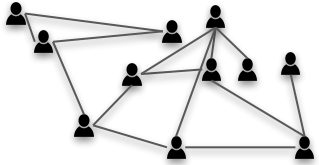
Generates the day of a user who walks nearby his home or hotel and checks into POIs.



The configuration parameters control:

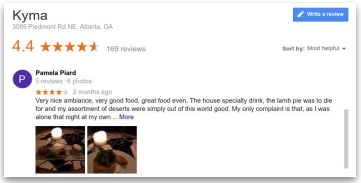
- how many daily routes?
- when does the day start and end?
- how many check-ins in a day?
- how long will a check-in last?
- how far can the user walk?

Output Data



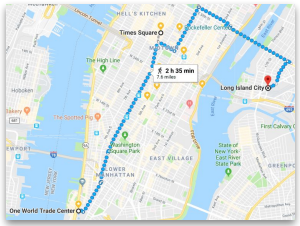
Social network

User → User



check-ins

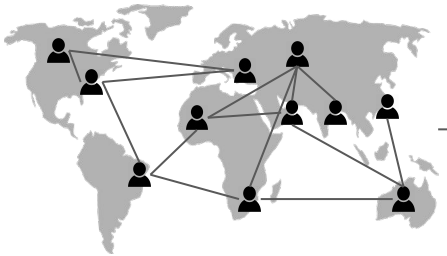
User → Check-in
- POI
- Review
- Time - Date



GPS traces

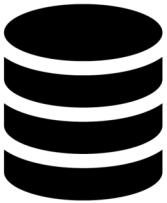
User → GPS Trace
- Latitude
- Longitude
- Time - Date

Storage - Queries



Geo-Social Network

Indexed by "user"



Database

Queries

For a random user:

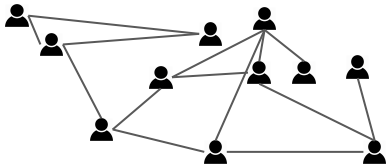
News Feed: Show all friend check-ins in chronological order.

What are the most favorite places that his friends have visited?

How many times have his friends been to their most favorite place?

?

Use Case

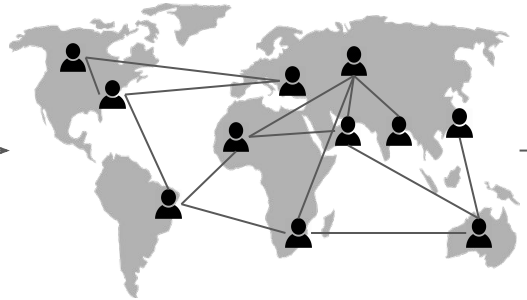


Twitter Graph = 14 GB

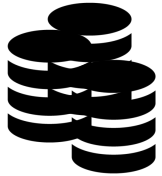
TripAdvisor restaurants = 13 GB

2 months
9 am - 11 pm
~5 check-ins / day
~2 hours / check-in
<0.5 miles between

Spaten



Geo-Social Network
14 + 3 = 17 GB
~10,000 users
(limited us of Google Maps API)



HBase cluster
32 nodes

Concurrent Queries

Summary

Code: https://github.com/Thaleia-DimitraDoudali/Spaten
Dataset: http://research.cslab.ece.ntua.gr/datasets/ikons/Spaten/