Σχεδιασμός και Υλοποίηση Γεννήτριας Χωροχρονικών Δεδομένων Μεγάλου Όγκου για Αποτίμηση Υπηρεσιών Κοινωνικής Δικτύωσης

Διπλωματική Εργασία

Θάλεια-Δήμητρα Δούδαλη

National Technical University of Athens

CSLab

Thesis contribution

1. Design and implementation of a parameterized generator of spatio-temporal and textual social media data
2. Creation of a large dataset using the generator
3. Storage of the dataset into an Hbase distributed database system
4. Scalability testing of the Hbase cluster

Motivation

- Era of Big Data
- Polymorphic social media data
- Transition to distributed storage and processing tools
- Limited access to such data due to privacy restrictions
- Restricted evaluation of distributed data management tools

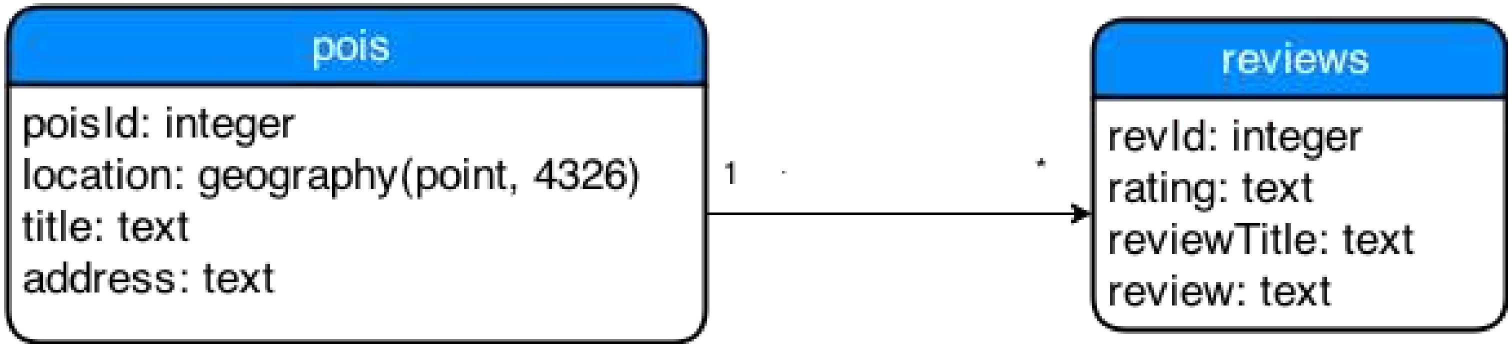
Generator

- Spatio-temporal and textual data
- Users of social networking service
- Daily Check-ins to Points of Interest leaving a review and rating
- GPS traces indicating the routes
- Static Map representation

Source Data

- Real Points of Interest crawled from TripAdvisor
- 136409 points = 13 GB JSON file
- Storage in PostgreSQL
- PostGIS extension offers functions and indexes for geographic data types

Source data schema



pois

poisId: integer
location: geography(point, 4326)
title: text
address: text

reviews

revId: integer
rating: text
reviewTitle: text
review: text

Input Parameters

- userIdStart, userIdEnd
- startTime, endTime
- startDate, endDate
- dist, maxDist
- chkNumMean, chkNumStDev
- chkDurMean, chkDurDev

Implementation

Check-ins:

- Number of daily check-ins defined using a gauss distribution
- First ever check-in = home location
- First check-in randomly chosen using uniform distribution
- It should be in maxDist range from home
- Rest check-ins of the day should be in walking distance (parameter dist)
- Assign random rating and review using uniform distribution

Implementation

Path between check-ins:

- Google Directions API
- JSON response file containing the path and duration
- Encoded polyline representation of the path
- Extracted geographical points as GPS traces

Implementation

Timestamps:

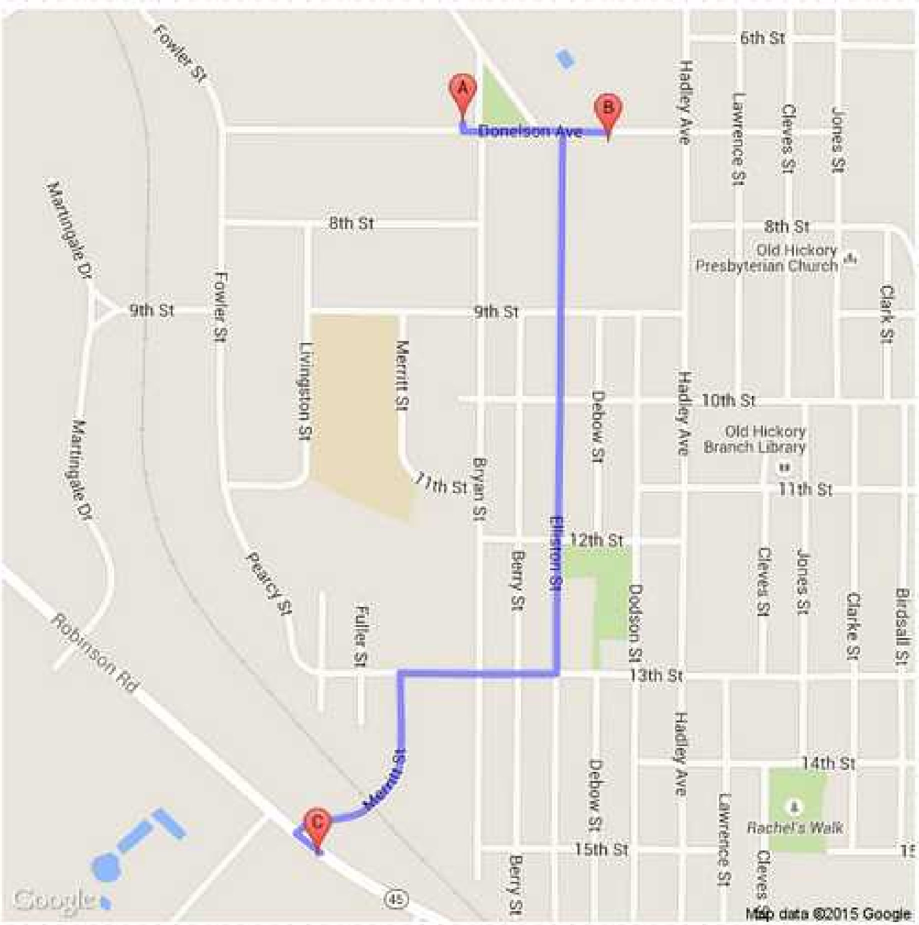
- First check-in of the day → startTime
- Duration of each visit → Gauss distribution
- Time of next check-in = time of previous one + duration of visit + duration of walk
- Should not exceed endTime
- GPS trace timestamp = splitted walk duration

Implementation

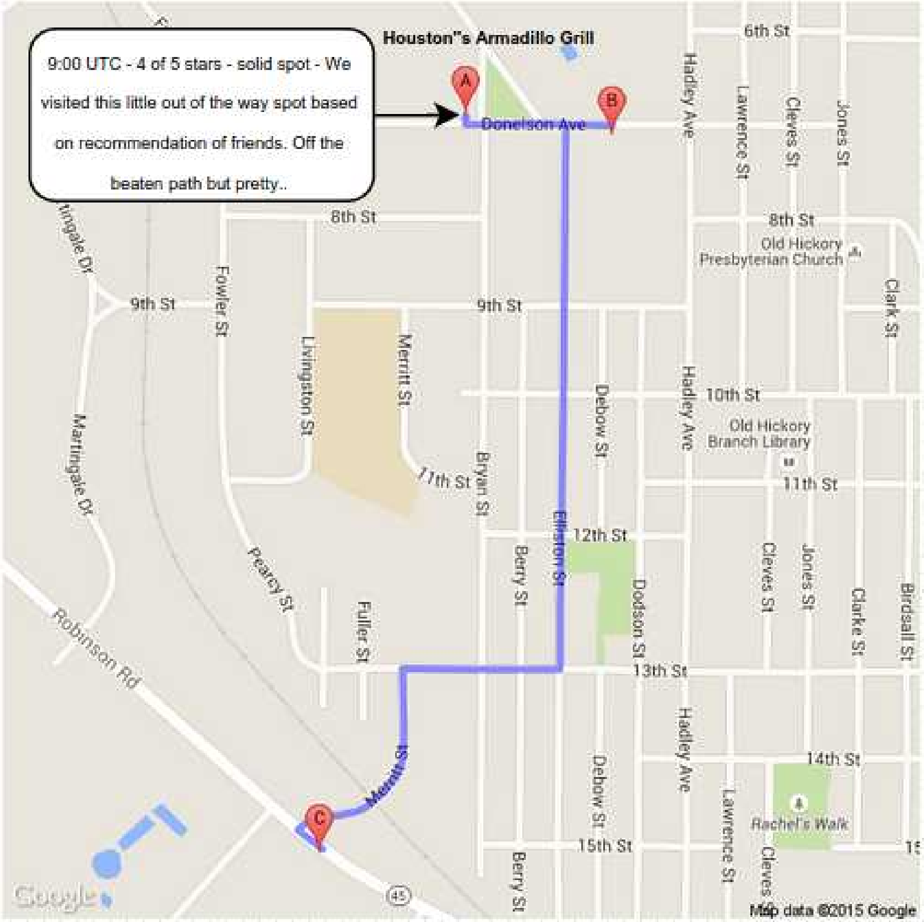
Trips:

- Travel location equivalent to home
- Available travel days = 10% (endDate – startDate)
- Trip duration = Gauss with $\mu = 5$ and $\sigma = 2$
- Decision to start trip → coin toss every day

Static Map

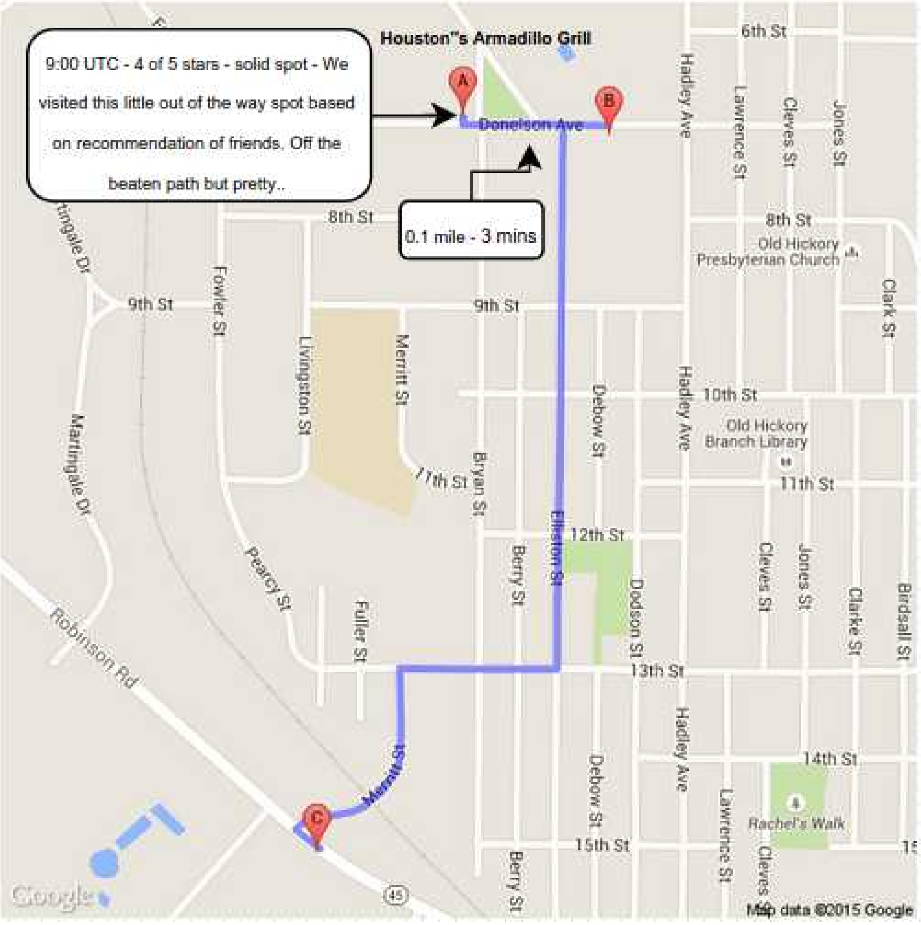

A
B
C
Donelson Ave
Elliston St
Merritt St
6th St
8th St
9th St
10th St
11th St
12th St
13th St
14th St
15th St
Fowler St
Martingale Dr
Livingston St
Bryan St
Debow St
Hadley Ave
Lawrence St
Cleves St
Jones St
Clark St
Clarke St
Birdsall St
Dodson St
Berry St
Fuller St
Pearcy St
Robinson Rd
Old Hickory Presbyterian Church
Old Hickory Branch Library
Rachel's Walk
45
Google
Map data ©2015 Google

Static Map



9:00 UTC - 4 of 5 stars - solid spot - We visited this little out of the way spot based on recommendation of friends. Off the beaten path but pretty..

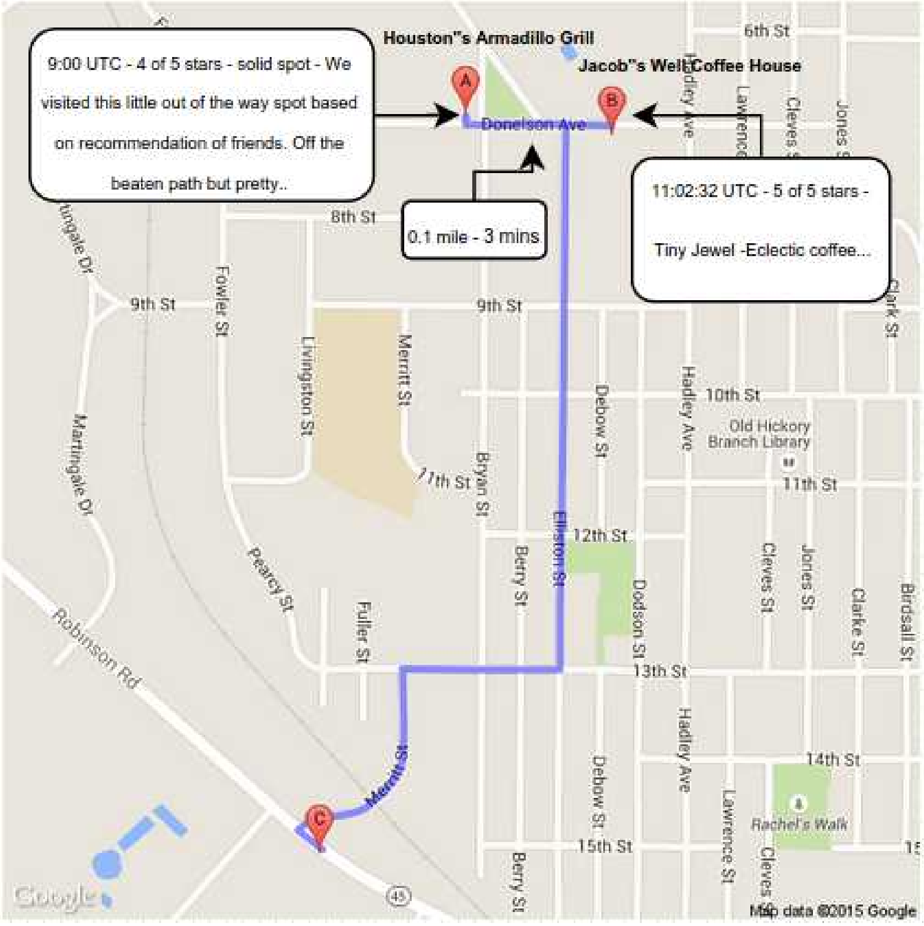
Static Map



9:00 UTC - 4 of 5 stars - solid spot - We visited this little out of the way spot based on recommendation of friends. Off the beaten path but pretty..

0.1 mile - 3 mins

Static Map

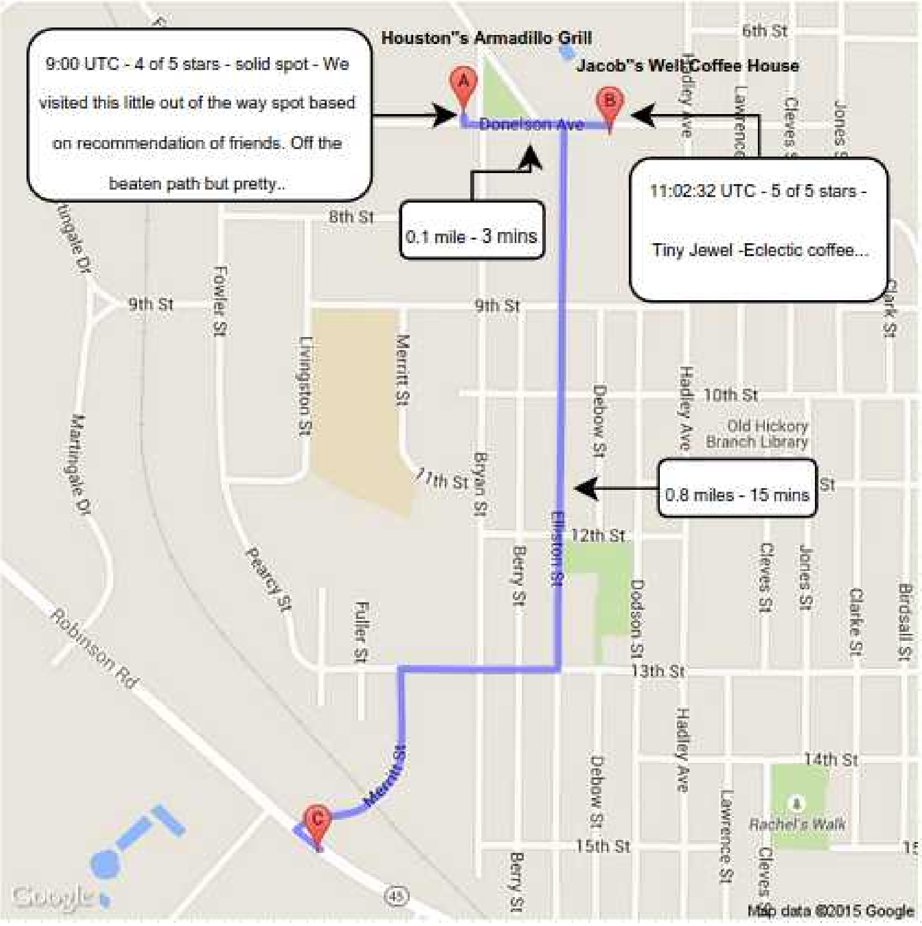


9:00 UTC - 4 of 5 stars - solid spot - We visited this little out of the way spot based on recommendation of friends. Off the beaten path but pretty..

0.1 mile - 3 mins

11:02:32 UTC - 5 of 5 stars - Tiny Jewel -Eclectic coffee...

Static Map



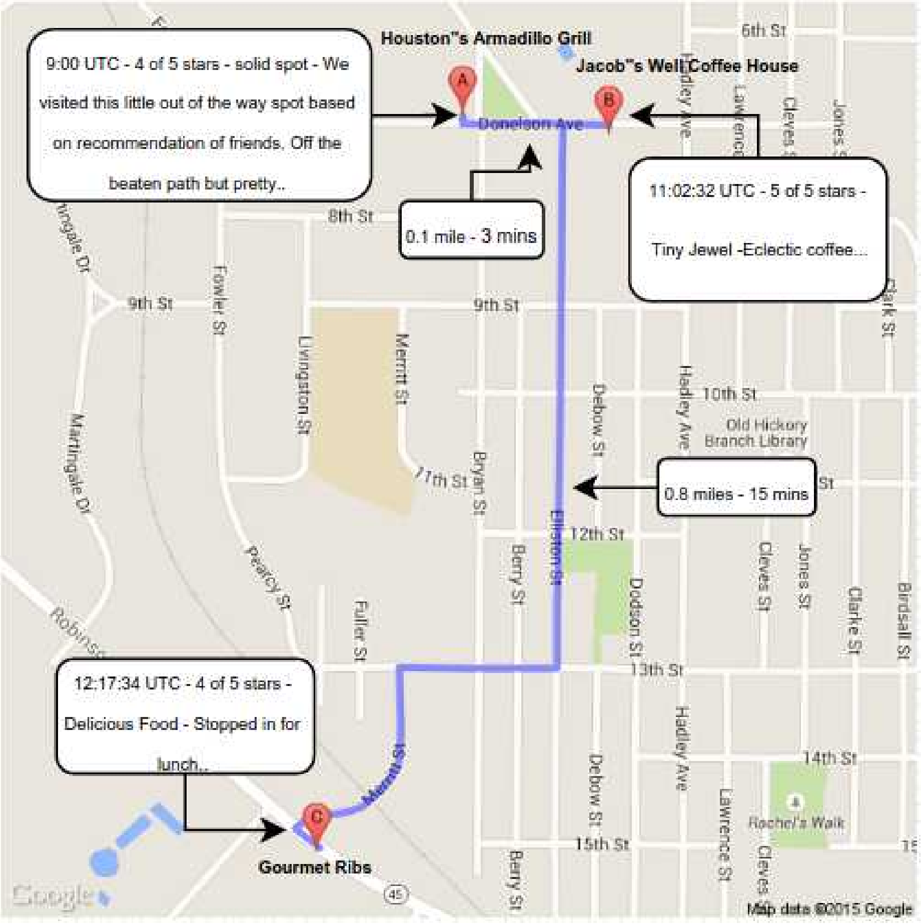
9:00 UTC - 4 of 5 stars - solid spot - We visited this little out of the way spot based on recommendation of friends. Off the beaten path but pretty..

0.1 mile - 3 mins

11:02:32 UTC - 5 of 5 stars - Tiny Jewel -Eclectic coffee...

0.8 miles - 15 mins

Static Map



9:00 UTC - 4 of 5 stars - solid spot - We visited this little out of the way spot based on recommendation of friends. Off the beaten path but pretty..

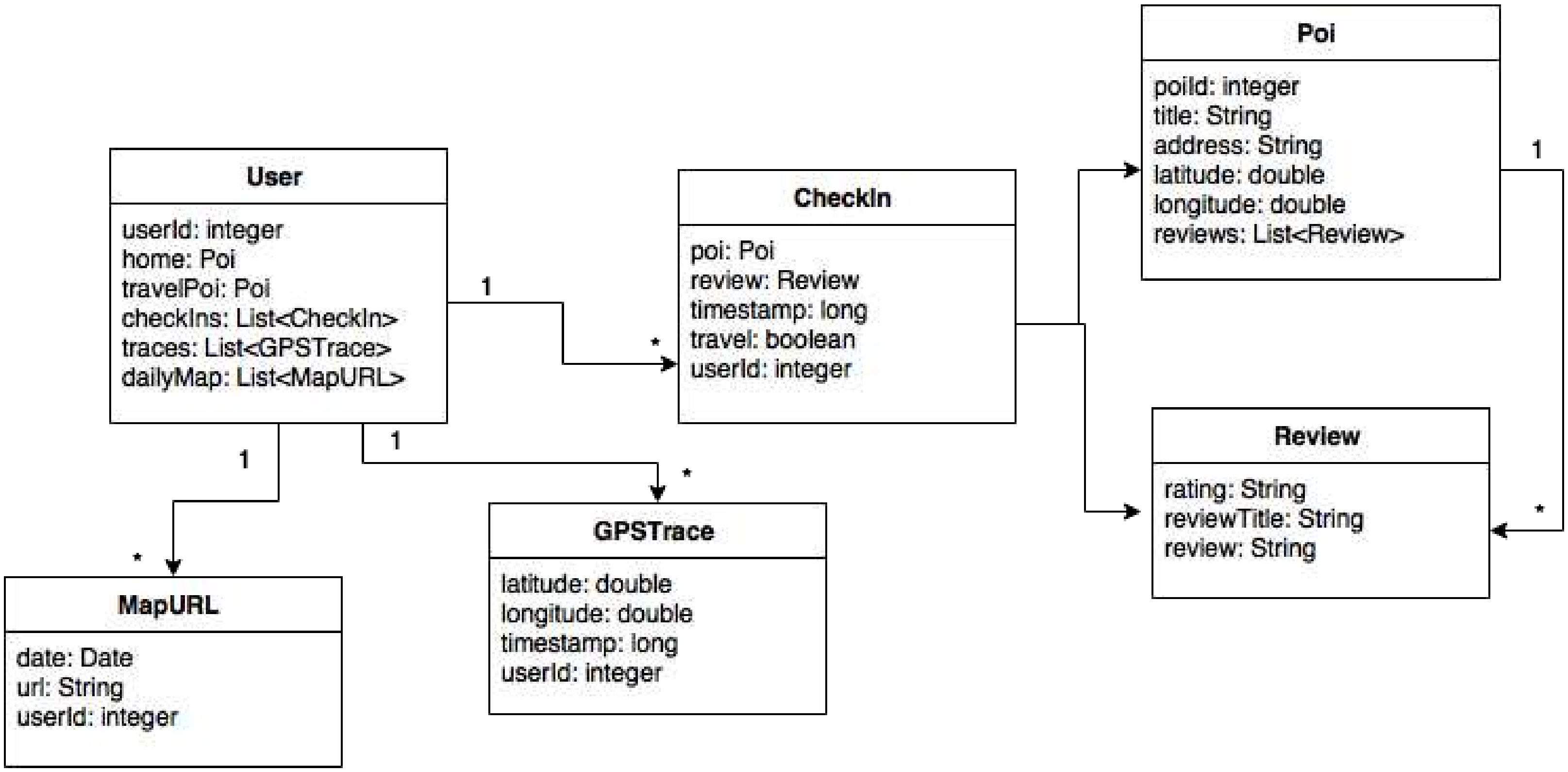
0.1 mile - 3 mins

11:02:32 UTC - 5 of 5 stars - Tiny Jewel -Eclectic coffee...

0.8 miles - 15 mins

12:17:34 UTC - 4 of 5 stars - Delicious Food - Stopped in for lunch..

Generator Attributes



User

userId: integer
home: Poi
travelPoi: Poi
checkIns: List<CheckIn>
traces: List<GPSTrace>
dailyMap: List<MapURL>

CheckIn

poi: Poi
review: Review
timestamp: long
travel: boolean
userId: integer

Poi

poiId: integer
title: String
address: String
latitude: double
longitude: double
reviews: List<Review>

Review

rating: String
reviewTitle: String
review: String

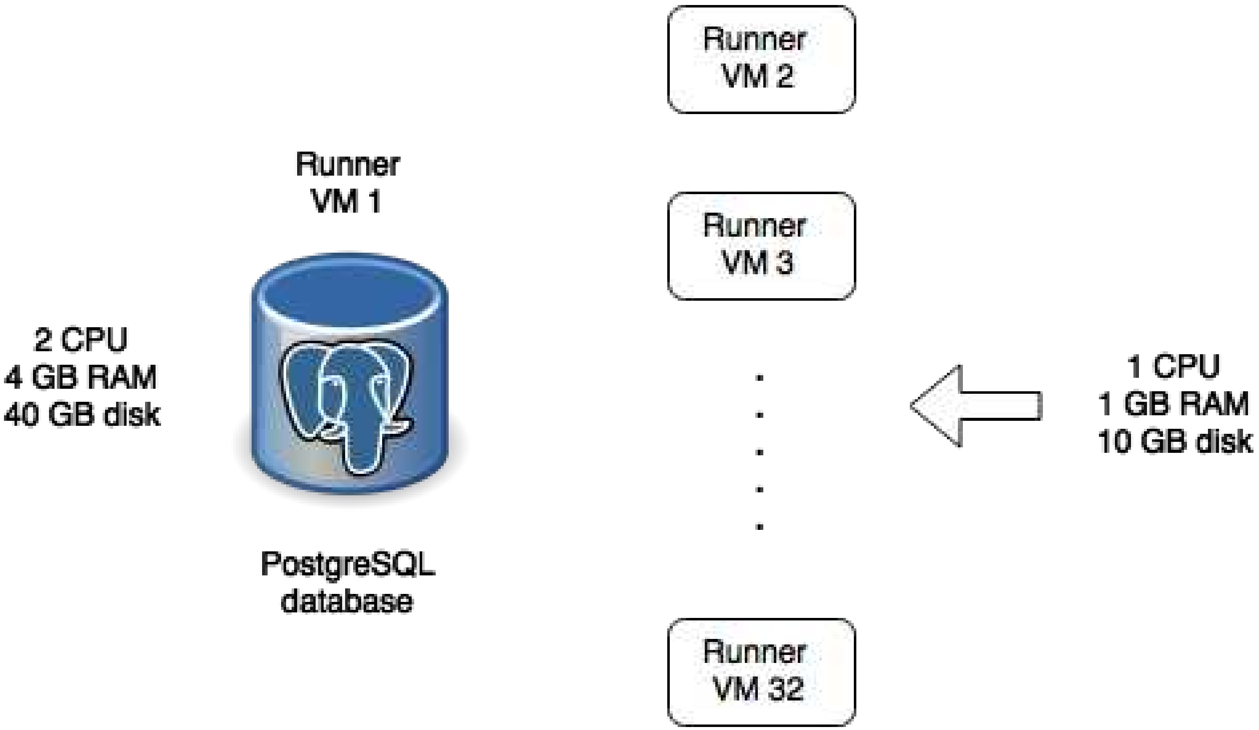
GPSTrace

latitude: double
longitude: double
timestamp: long
userId: integer

MapURL

date: Date
url: String
userId: integer

Generator Deployment Setup



Runner
VM 1

2 CPU
4 GB RAM
40 GB disk

PostgreSQL
database

Runner
VM 2

Runner
VM 3

Runner
VM 32

1 CPU
1 GB RAM
10 GB disk

Execution Input Parameters

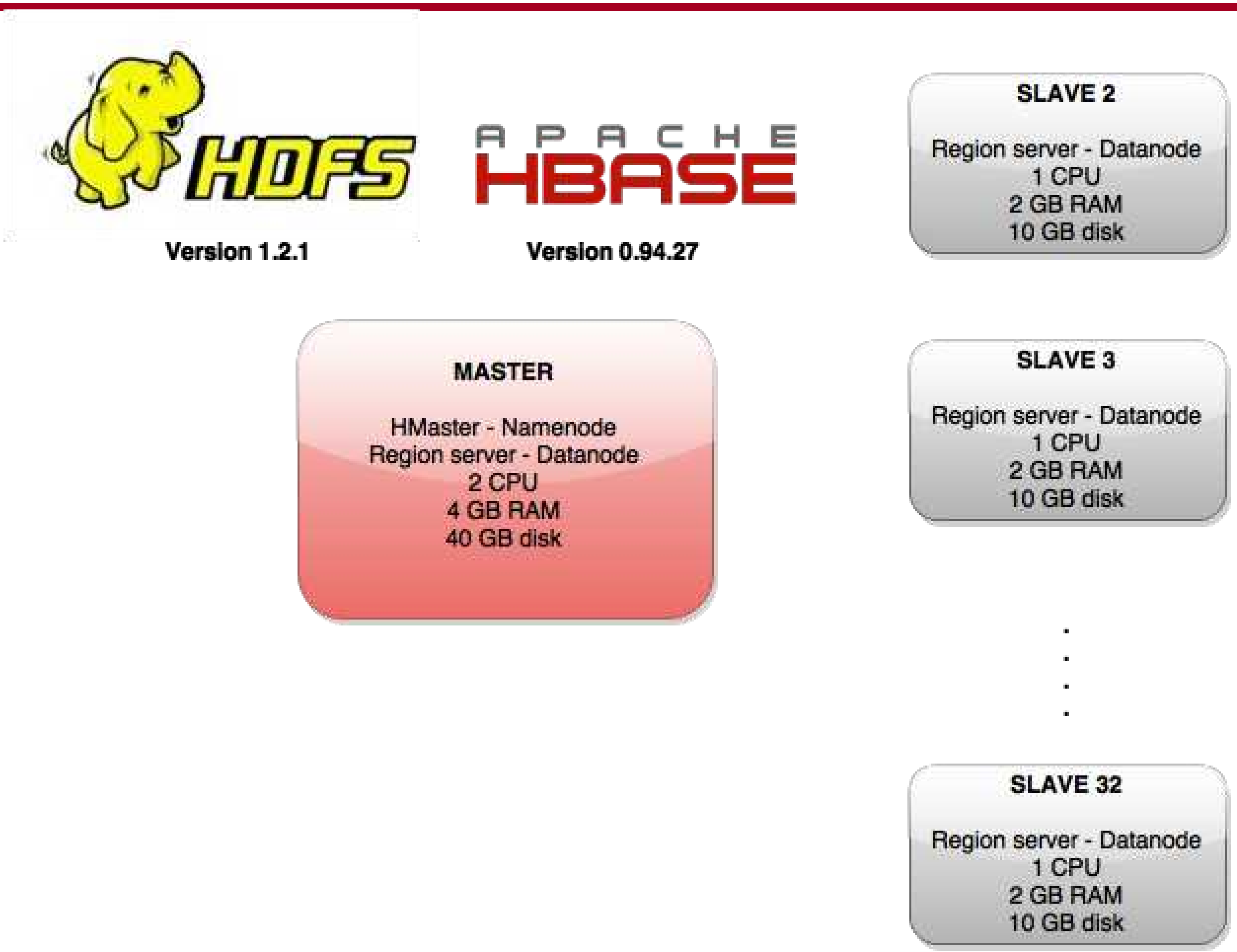
- chkNumMean = 5 chkNumStDev = 2
- chkDurMean = 2 chkDurStDev = 0.1
- maxDist = 50000.0 dist = 500.0
- startTime = 9 endTime = 23
- startDate = 01-01-2015 endDate = 03-01-2015

Generated Dataset

- 9464 users with 2 months daily routes
- 1,586,537 check-ins → 641 MB
- 38,800,019 GPS traces → 2.4 GB

- Added a 14 GB twitter friend graph

HBase cluster



Version 1.2.1

Version 0.94.27

MASTER

HMaster - Namenode
Region server - Datanode
2 CPU
4 GB RAM
40 GB disk

SLAVE 2

Region server - Datanode
1 CPU
2 GB RAM
10 GB disk

SLAVE 3

Region server - Datanode
1 CPU
2 GB RAM
10 GB disk

.
.
.
.

SLAVE 32

Region server - Datanode
1 CPU
2 GB RAM
10 GB disk

HBase data model

- Friends table
  - Row: user id
  - Column Qualifier: friend user id
  - Cell Value: friend user id

- Check-ins table
  - Row: user id
  - Column Qualifier: timestamp
  - Cell Value: check-in data

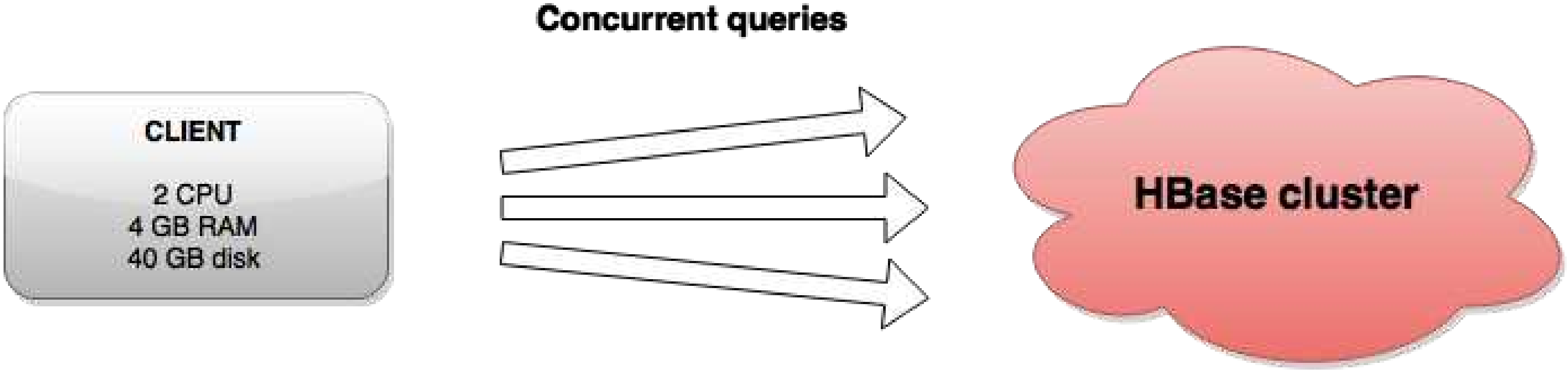
- GPS traces table’
  - Row: user id
  - Column Qualifier: “lat long timestamp”
  - Cell Value: GPS trace data

Queries

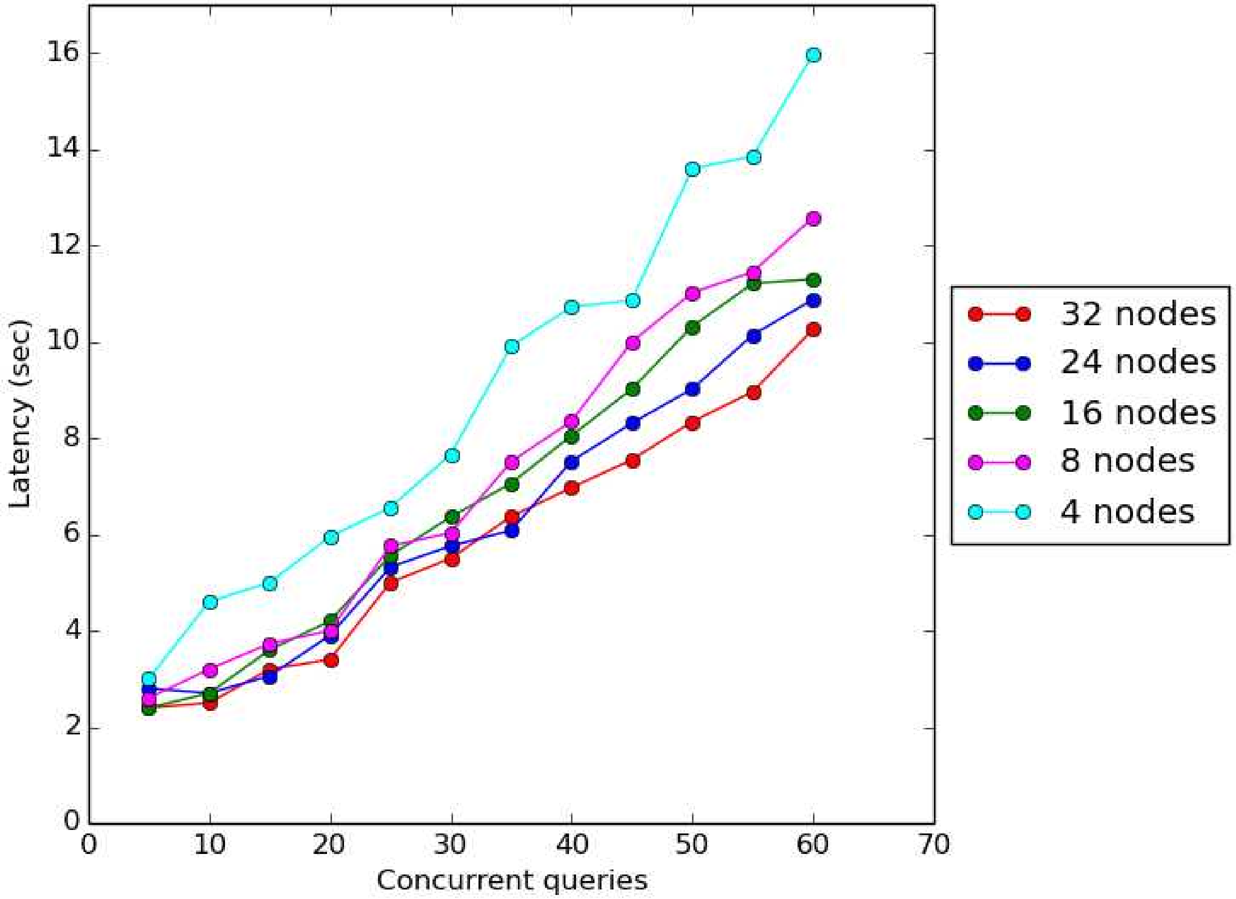
1. Get the most visited points of interest of a certain user’s friends
2. Get the check-ins of all the friends of a specific user for a certain day into chronological order (News Feed)
3. Get the number of times that a user’s friends have visited the user’s most visited POI

Implemented using HBase coprocessors on data balanced region servers

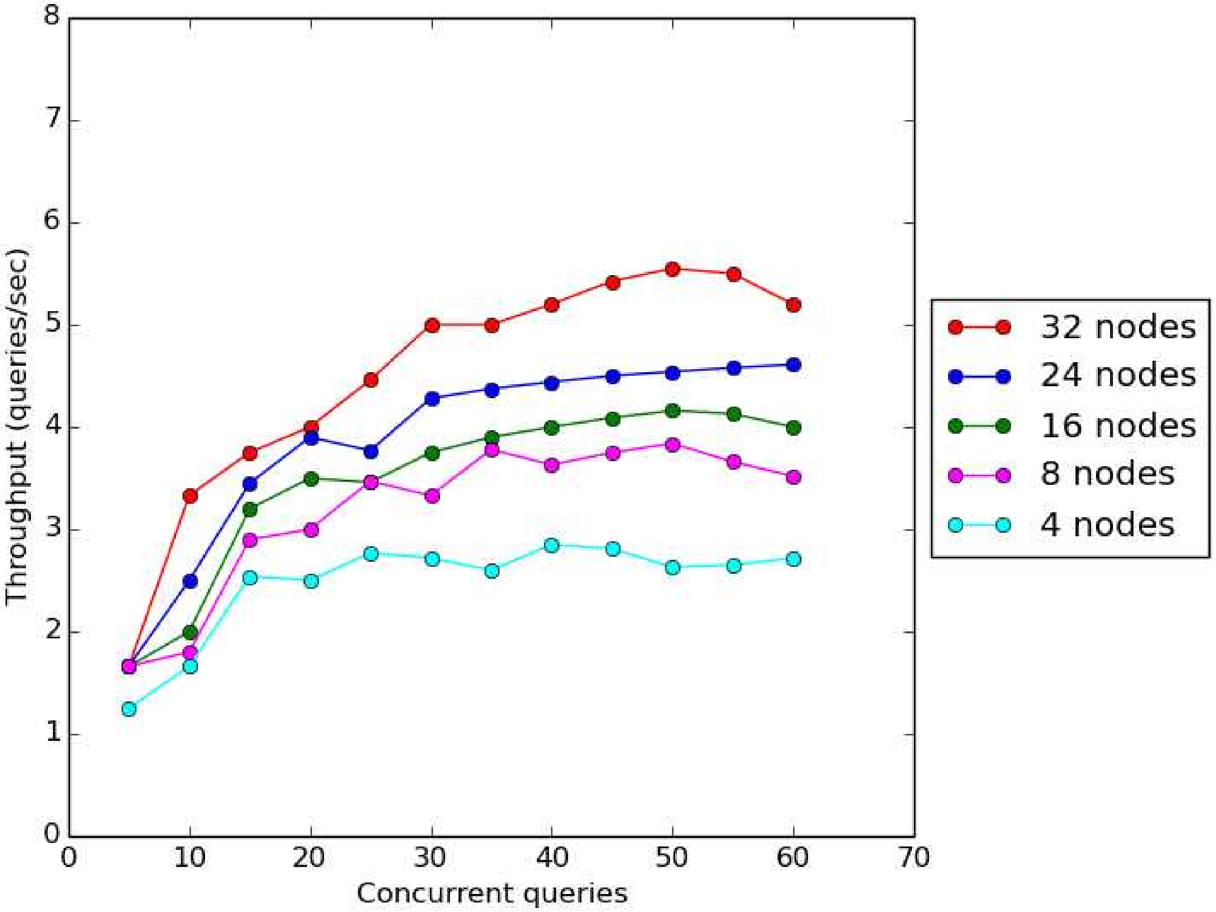
Workload generation setup



Scalability Testing



Scalability Testing



Conclusion

- HBase cluster is scalable for the specific data storage model of the dataset produced by the generator
- HBase provides indeed good performance and data management tools for Big Data social networking services

Questions

?